CONSTRUCCIÓN DEL ROL DE LAS MUJERES DEL SIGLO XIV Y XV EN PLASENCIA ESPAÑA, A PARTIR DE DOCUMENTOS HISTÓRICOS

APLICACIÓN DE MINERÍA DE TEXTO A DOCUMENTOS HISTÓRICOS

Noviembre de 2017
Universidad Tecnológica de Pereira
facultad de ingeniería industrial.

APLICACIÓN DE MINERÍA DE TEXTO A DOCUMENTOS HISTÓRICOS:

CONSTRUCCIÓN DEL ROL DE LAS MUJERES DEL SIGLO XIV Y XV EN PLASENCIA-ESPAÑA, A PARTIR DE DOCUMENTOS HISTÓRICOS.

Estudiante: Piedad Catalina González Orjuela

Trabajo de grado para optar por el Título de:

Magister en Investigación Operativa y Estadística.

Director: Dr. Ignacio Pérez

Co – director: Dr. Roger Martínez

Universidad Tecnológica de Pereira UTP

Facultad de ingeniería industrial

Pereira

2017

AGRADECIMIENTOS

Dr. Ignacio Pérez & Dr. Roger Martínez, por proveerme las información y asesoría necesaria para la ejecución del proyecto de investigación.

A mi familia por su apoyo.

Tabla de **contenido**

1. INTRODUCCIÓN.

La historia de la humanidad es parte fundamental en el entendimiento de las dinámicas sociales actuales, en tanto permite conocer cómo se llegó a las relaciones humanas, en aspectos: religiosos, económicos, gubernamentales y eclesiásticos; son los documentos históricos una fuente de recopilación de información esencial para este entendimiento, ya que permiten evidenciar transacciones comerciales, interacción entre religiones, organizaciones jerárquicas, entre otros aspectos.

Los documentos históricos son una fuente de información extensa que vale la pena analizar. Más allá de una lectura superficial, a lo largo del tiempo se ha hecho un mayor énfasis en el análisis de datos numéricos y el desarrollo de métodos estadísticos con datos cuantitativos, sin embargo, a partir de la década de los noventa se volcó la atención al análisis de textos, por contener información valiosa, así que, en la conferencia de inteligencia Artificial de agosto de 1999, apareció por vez primera el termino minería textual (Witten & Frank, 2005).

La información textual tiene relaciones ocultas, esto quiere decir que es probable que existan variables correlacionadas, que no se perciben fácilmente, como sucede en ocasiones con los datos numéricos, de los cuales se ocupa la minería de datos cuantitativos; en el caso de textos el análisis de documentos se hace por minería de texto, la cual busca obtener información oculta en la redacción y estructura de la misma. No se debe confundir la minería textual con la indexación o recuperación de información (categorización y clasificación), sino que se trata de un análisis de los datos –profundo- que busca sacar relaciones escondidas y no obvias, esto mediante métodos estadísticos (Textmining.galeon.com, 2015)

Dada esta nueva herramienta estadística y computacional se genera un especial interés por el análisis de textos. El Dr. Roger Martínez, profesor de historia de la Universidad de Colorado, es especialista en: paleografía medieval española (Técnica que consiste en leer los documentos, inscripciones y textos antiguos y en determinar el lugar del que proceden y el período histórico en el que fueron escritos), historia entre las relaciones interreligiosas medievales y la historia de la comunidad de Plasencia España, lidera el proyecto de investigación “The Revealing Cooperation and Conflict Project”, que consiste en recrear, de

forma virtual, la ciudad de Plasencia del siglo XIV y XVI, *a través de un conjunto de documentos encontrados en la catedral de Plasencia*. El objetivo principal de su proyecto es reconstruir la interacción entre los judíos, los cristianos y los musulmanes, razón por la cual, se recrearán los procesos de cooperación y también algunas de las disputas que surgieron durante ese periodo de violencia, esto se realiza a partir de la recolección de ***actas/ manuscritos*** de clérigos católicos, familias judías de nobles, mercaderes, señores medievales y de clanes de caballeros.

Estos documentos son manuscritos en español antiguo como se muestra en la figura 1, su legibilidad y español no es cotidiano, por lo que el Dr. Martínez crea un curso de Paleografía en el portal Coursera, en dónde da instrucciones de cómo realizar las transcripciones por parte de los estudiantes, dando segmentos similares a los de la fig.1 y pide que los estudiantes descifren lo que allí esta escrito y transcriban estos a medios electrónicos para posteriormente se envíen al equipo de profesores del curso. Estas transcripciones se consolidaron en archivos en Excel y páginas HTML, acompañadas con las observaciones del equipo de investigadores con las apreciaciones de los evaluadores de las transcripciones elaboradas.

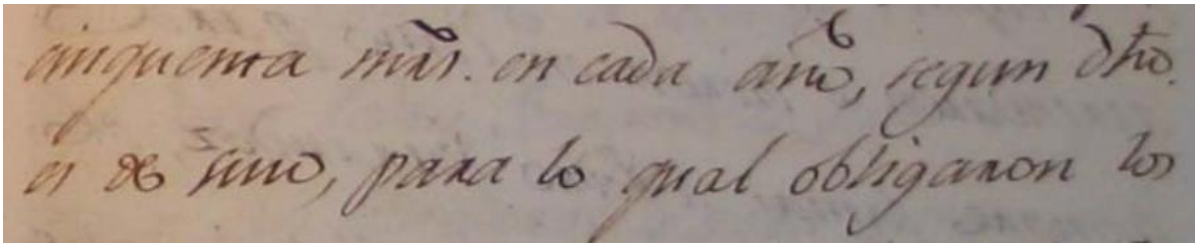


Fig. 1 Ejemplo de acta capitular. (Puglisi & Martinez, 2014).

Se tiene un volumen de 10125 transcripciones, de diversos fragmentos de las actas, dado que a diferentes grupos de estudiantes se les entregó el mismo fragmento, por lo que para un mismo fragmento existen diferentes versiones de la transcripción, por lo que la reconstrucción de los manuscritos es un poco compleja de forma manual.

A partir de este proyecto, el Dr. Ignacio Pérez, cercano a la investigación del Dr. Matinez, ve el problema de reconstrucción de Plasencia, una oportunidad de implementar usando minería textual en las actas capitulares, ya que se está ante un volumen de datos considerable y el campo textual ha sido poco explorado.

A partir del interés de aplicar minería en la optimización de procesamiento y entendimiento de grandes volúmenes de datos, se plantea un proyecto de investigación implementando un ejemplo de análisis de minería textual a los Documentos de la Catedral de Plasencia, el objetivo del presente estudio es construir un prototipo de ejemplo funcional de minería de texto, mediante la búsqueda del papel de las mujeres dentro de los documentos del municipio de Plasencia ubicado en la provincia de Cáceres. Este prototipo funcional puede ser utilizado posteriormente por la comunidad de historiadores para tareas similares.

1.1. Planteamiento del problema.

Para la selección del tema a trabajar en el prototipo se consideró que el énfasis de la investigación del Dr. Martínez, abarca aspectos religiosos y económicos de los habitantes de Plasencia, se detectó que no era un objeto de estudio la caracterización de la población de la provincia, por lo que aspectos personales no eran primordiales, en la investigación.

Se hace una lectura rápida de algunos fragmentos y es curioso no encontrar mención alguna de la participación de las mujeres de la época y aunque la historia conocida de estos siglos, nos habla de la poca participación de las mujeres en asuntos comerciales, era importante para la investigación, reafirmar esto o descubrir otro tipo de correlaciones dentro de los manuscritos referentes a la mujer. Dada la riqueza histórica de los documentos es una oportunidad de obtener más información al respecto de este tema.

De la iglesia, de hombres influyentes de la época, de las interacciones comerciales y religiosas, era ya objeto de estudio de los investigadores del Dr. Martínez, por lo que además de hacer un aporte desde el análisis de textos a través de la Minería, se pensó en hacer un aporte en sentido estrictamente histórico, no mencionado de las dinámicas de Plasencia de los siglos XIV y XV, al estudiar *el rol de las mujeres de la época*, a partir de la Minería Textual, siendo la mujer parte esencial en el desarrollo y evolución de las comunidades, es importante conocer las posiciones ocupadas en la provincia.

Sin ser la autora historiadora, como se mencionó antes se quiere poner al servicio de esta disciplina un proceso que permite hacer el análisis de cualquier conjunto de textos históricos,

escritos en español antiguo, dicho proceso permite hacer un análisis de los temas que aparecen en los textos, así como de encontrar correlaciones entre términos, *sin necesidad de hacer una lectura previa de los textos*. Cabe repetir que el proceso planteado en esta investigación se puede implementar en cualquier tema que sea objeto de estudio, como se verá en las secciones siguientes.

El Dr. Martínez, se encarga de hacer un análisis de estos manuscritos desde el análisis histórico de texto, que consiste en seguir los pasos que se explican a continuación (Comentario de textos, 2017):

1. *Lectura y preparación*: Se divide en dos lecturas preliminares, la primera sin extracción de información, luego una lectura comprensible, en dónde se subrayan términos relevantes, ideas primarias, ideas secundarias y se realizan anotaciones marginales.
2. *Clasificación del texto:* se debe especificar la naturaleza del documento, las circunstancias espacio. temporales, autor y destino del mismo.
3. *Análisis del texto:* este se basa en, el método literal: que consiste en seguir un orden descriptivo con la explicación progresiva de palabras, expresiones y alusiones que en él aparecen; El método lógico: consiste en reagrupar los pasajes y las explicaciones de acuerdo con su temática; resulta útil para textos mal articulados o confuso.
4. *Comentario del texto:* se trata de tomar el texto como fundamento o base para desarrollar y disertar sobre el momento histórico en sus aspectos más generales.
5. *Crítica del texto:* determina su autenticidad y exactitud, su sinceridad y objetividad, el interés, es decir si se trata de un documento decisivo para el análisis del momento histórico en el que se inscribe o hace referencia, o por el contrario tiene una importancia relativa o secundaria.
6. *Bibliografía*: Además de las fuentes consultadas directamente durante el análisis del texto es aconsejable, en la manera de lo posible, hacer mención de: Manuales y obras generales, Libros especializados y monografías entre otros.

Este es un trabajo dispendioso que debe llevar a cabo el historiador, que en algunas ocasiones presenta los siguientes problemas: **Digresión o disertación** que implica utilizar el texto como

simple pretexto para exponer los conocimientos del historiador sobre algún tema, puede llevar a desviarse del asunto central, **Paráfrasis,** es exposición repetitiva de lo que dice el texto, sin aportar medios para su interpretación y **Personalismo:** que consiste en expresar opiniones o juicios desde un punto de vista personal. Problema que sería inadmisible, puesto que el ejercicio de ***Historia ha de perseguir la objetividad y la total ausencia de prejuicios*** (Comentario de textos, 2017)

En aras de minimizar estos problemas y lograr la objetividad del análisis de textos históricos, el uso de la técnica de Minería Textual para hacer análisis de textos históricos desde una perspectiva matemática, en dónde el historiador con este proceso no tiene la necesidad de hacer una lectura previa, para encontrar los temas que se están tratando en el texto, es decir la parte de clasificación de términos y de temas con el proceso que se implementó, se optimizan ya que se hace de forma automática. Adicionalmente brinda la posibilidad de hacer extracción de nombres de lugares y personajes que aparecen en el texto lo cual complementa y agiliza el desarrollo de las investigaciones históricas con la metodología de esta disciplina explicado arriba.

Una ventaja del análisis automático de texto es que permite entender información no estructurada expresada en una lengua y la convierte en información estructurada, esto puede ser como el resumen de su contenido, relacionando los elementos más significativos o clasificando un documento por su tipología (El análisis automático de texto con Big Data e Inteligencia Artificial - IIC, 2017).

1.2. Objetivos.

El objetivo principal de esta investigación es desarrollar un prototipo de tareas de Minería Textual que permite hacer un análisis de Textos Históricos, en español antiguo, el objeto fundamental es mostrar un ejemplo del potencial de la técnica buscando conocer el papel que desempeño las mujeres en los siglos XIV y XV, dada la **no** mención en el trabajo investigativo del Dr. Martínez.

El desarrollo de este proceso de Minería Textual, busca proveer una nueva metodología para el análisis de textos históricos, que pueda ser de fácil acceso para los historiadores, se propone implementar un software amigable para estos investigadores que en general no están familiarizados con lenguajes computacionales complejos. Uno de los programas que permite una interacción sencilla es KNIME, este ofrece un código en el que visualmente se entiende el proceso de Minería Textual de los documentos, sin requerir un conocimiento profundo de lenguaje computacional especializado. Este es uno de los factores que hace que KNIME sea elegido para este proyecto, además de brindar diferentes opciones de minería de texto, con ejemplos prácticos de implementación para no expertos en programas de analítica.

Adicionalmente, la creación de un proceso automático constituye la propuesta de este trabajo, basada en un clasificador de temas, que busca estructurar y representar el contenido conceptual de textos, a través del método de asignación latente de Dirichlett (Latent Dirichlet Allocation LDA). Este método parte de la idea básica que los documentos a analizar son una mezcla de temas, donde cada tema es una variable la cual se caracteriza por una mezcla de palabras. Así el proceso a implementar permitirá conocer los temas que en los manuscritos se hablan, buscando especialmente si en estos se encuentra la participación de la mujer.

Mostrar el análisis arrojado del procesamiento de los documentos también es uno de los objetivos del trabajo, lograr obtener temas de los manuscritos sin leerlos, ni ser especialistas en historia.

2. FUNDAMENTACIÓN TEÓRICA.

2.1. Minería de Texto.

El tratamiento de datos cuantitativos es más común en las ciencias, la estadística, la economía, la física entre otras ramas, estas se encargan del análisis de resultados numéricos de experimentos en cada área, la estadística y probabilidad han desarrollado técnicas para la predicción, modelamiento, e inferencia a partir de datos tomado de situaciones particulares (Estadística para todos, 2017), sin embargo existe un gran volumen de datos en forma de texto a los cuales por métodos estadísticos tradicionales es difícil analizarlos, por lo que la Minería de Datos, creo una rama la cuál denominó *minería de texto*, esta rama es concebida como el tratamiento de datos de tipo cualitativo, como: documentos, libros, descripciones médicas, artículos científicos, artículos de periódicos, documentos históricos. Su finalidad es obtener información *oculta (correlaciones no obvias entre temas o palabras)* en la redacción y estructura de un conjunto de documentos.

Esto no se debe confundir esto con la indexación o recuperación de información (categorización y clasificación), ya que la *minería de texto* es un análisis de los datos –profundo- que busca sacar relaciones escondidas, que no son obvias, como se afirma en la página web de *textmining.galeon.com* (2015), es hacer un esfuerzo por ir más allá de la lectura de documentos y su análisis, ya que la idea central es aplicar técnicas probabilísticas sobre las palabras de cada texto.

La importancia de la *minería de texto* permite analizar gran extensión de lenguaje textual, detectando el léxico o patrones lingüísticos, está dentro de los métodos de análisis del *Big Data* (Witten & Frank, 2005).

Como toda actividad investigativa la *minería de texto* se tienen tres etapas fundamentales para su implementación (cfr. Textmining.galeon.com, 2015).

- Determinación de los objetivos.
- Pre procesamiento de los datos: selección, análisis y reducción de los textos.
- Determinación de la técnica o modelo a seguir.

- Análisis de resultados. Hacer conclusiones que permitan la toma de decisiones.

De esta forma el objetivo será facilitar la toma de decisiones, así como el entendimiento del texto que se estudia (Textmining.galeon.com, 2015).

En el software que se desarrolló el código principal de la investigación Knime, muestra la importancia y finalidad de cada uno de los pasos anteriormente mencionados, se muestra y se explica cómo se llevó a cabo en la investigación (KNIME, 2013):

- Lo primero que se debe hacer es *establecer el formato de los documentos* según el software que se implemente; el software a utilizar en este proyecto es KNIME, el nodo lector de documentos XML, es un formato de texto que permite, procesar archivos Xls, dado que los datos están en hojas de Excel es óptimo para la lectura de las transcripciones.
- El siguiente paso tiene que ver con la *preparación de los datos*; se debe construir el corpus del documento, el cual se define como un conjunto de textos con criterios lingüísticos explícitos para asegurar que pueda usarse como muestra representativa de una lengua (cfr. Elies.rediris.es, 2015), es decir, el conjunto de todos los documentos, consolidando los datos en una única estructura., el Corpus: es matemáticamente una colección de M documentos denotada por M={ w1, w2,…., wM }, el corpus de los manuscritos se disponen en archivos xls.
- *Limpieza de los datos*. En este paso se debe filtrar la información para dejar los datos más relevantes dentro del texto para evitar inconvenientes con las frecuencias calculadas. Se debió quitar caracteres como: artículos, preposiciones, números, remover la puntuación, conectores y pronombres, las cuales son conocidas como *Stop Words o palabras vacías*, se hace necesario omitirlas ya que estas palabras no aportan para la relación de términos dentro del documento.
- A continuación, se hace el *análisis morfológico* de las palabras, las oraciones, los párrafos y, finalmente, los textos. Para esto, inicialmente se aplica la "*tokenización*", que consiste en unir palabras con un tema común, y a éstas, unificarlas por una palabra, por ejemplo, Mahoma, Isaías, Joel, son palabras que se refieran solo a

profetas, lo que permite crear un token, el cual sería “profetas”. Esta práctica ayuda a reducir la dimensionalidad de los textos a hacer analizados, ya que compacta información de los párrafos.

- *Reemplazar*. Se debe filtrar una única palabra y quitar los sinónimos, y así unificar información, por ejemplo, si habla de niña, dama, señorita, señora, se puede reemplazar, simplemente, como mujer, para evitar un sesgo en el cálculo de la frecuencia de las palabras, ya que puede haber doce veces mencionada niña, tres veces dama, etcétera, en vez de esto, se calcula una única vez con la palabra mujer.
- *Stemming (derivación).* A los verbos se les debe quitar las conjugaciones y dejarlos en infinitivo, por ejemplo, si está jugando, debe quedar jugar, igualmente unifica al verbo general y evita errores en el cálculo de la frecuencia.
- Dentro del análisis, se debe extraer una bolsa de palabras (*bag of words*), que permitan etiquetarlas o clasificarlas. Este método también se puede emplear con frases relevantes para hacer una clasificación de las mismas, la intencionalidad es definir si es un verbo, sustantivo, nombre, lugar, el software trae definido un diccionario en español, sin embargo, se debe recordar que el español del cual se analiza es antiguo, lo que provoca que se cree un diccionario de español antiguo para este fin, se detallará en datos y métodos.
- *Cálculo de las frecuencias*. El cálculo de frecuencias es importante porque ayuda a caracterizar los documentos y sus relaciones, así se pueden determinar las palabras más o menos empleadas. La frecuencia de los términos se puede calcular como una frecuencia absoluta y relativa (# de ocurrencias del término dividido entre el número de términos). La relevancia de una palabra se puede ver bajo la idea de que la frecuencia de aparición de una palabra en un texto es inversamente proporcional, a la posición que ocupa en el ranking de frecuencias de palabras de un texto, esto es conocido como la Ley de Zipf (Moreiro González, 2002).

  Al ordenar las frecuencias de forma decreciente, se relaciona inversamente la posición de la frecuencia con, la frecuencia misma, de esta forma el aumento de rango o de posición implica una disminución en la frecuencia. Está ley de Zipf se aplica de acuerdo con la siguiente metodología (Moreiro González, 2002):

- Ordena de forma decreciente las frecuencias.
- Al multiplicar la frecuencia con la posición de aparición debe dar un valor aproximado a una constante.
- Obtención de la media para términos de frecuencias iguales, su efecto es la disposición en orden alfabético.
- Las palabras que tuviesen frecuencia superior a la constante C son elegidas.

Esta ley establece un umbral en dónde tiene en cuenta las frecuencias, si están muy por encima de la constante estas son palabras que aparecen mayormente pero no sirven para el análisis del texto, se consideran palabras vacías, si está muy por debajo son muy raras por lo que también se consideran vacías, las cuales deben ser eliminadas del documento. Sin embargo, el método es efectivo para altas frecuencias, para frecuencias bajas no funciona la misma constante, es el caso de artículos científicos donde el lenguaje es más técnico y limitado.

Una de las técnicas que permiten identificar los términos relevantes entre los de mayor y menor frecuencia para el análisis, es el método de indexación estadística de términos por frecuencias **IDF**, el cual es un sistema de filtrado basado en **Zipf**, estableciendo un sistema de pesos en función de la frecuencia relativa de cada término en cada documento, según este término (Moreiro González, 2002):

$$IDF = \log\left(\frac{1 + \#\ de\ documentos}{\#\ de\ documentos\ con\ el\ termino\ t}\right)$$

A partir de este peso fijado si es mayor a la media de los documentos, este será tomado como descriptor del resto de los documentos. De esta forma quedan únicamente los términos relevantes dentro de los documentos. Este método de frecuencias es alternativo al de frecuencias inmediatas que es el aplicado en la investigación.

Ahora se debe de *clasificar* los documentos, ya limpios y filtrados, para determinar las relaciones ocultas que se quieren identificar. Estos documentos se deben disponer como vectores, formando un espacio vectorial conocido como el espacio de los documentos o Tuplas. En este punto entra la minería de datos para la clasificación, la cual se puede hacer si se tienen algunas categorías por: árboles de decisiones, redes neuronales, etcétera; si no se

tienen las categorías como es el caso de esta investigación se recurre a *Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (pLSA), LDA Latent Dirichlet Allocation, y Non-negative Matrix Factorization (NMF).*

La minería textual implementada, por tanto, debe de considerar el tamaño de los datos. Según Charu, c, et al, (2012), la principal característica de los datos textuales es que son escasos y de alta dimensionalidad; estos problemas de dimensionalidad se han tratado con la *tokenizació*n y las *coocurrencias*, esta última, conocida como un indicador de la relación entre dos palabras cuando aparecen en un mismo párrafo.
A continuación, se caracterizarán los métodos de aprendizaje no supervisado, ya que estos métodos son los que competen a la investigación, dado que no se tienen las categorías a clasificar los temas de los documentos.

2.2. Clúster de documentos según Witten & Frank, (2005).

El Clúster de documentos es un tipo de *aprendizaje no supervisado*, es decir que la agrupación o clasificación no conoce de antemano las categorías a clasificar (Witten & Frank, 2005). Las frases dentro de los documentos son los datos representados como vectores, adicionalmente, por clase se escogen palabras y se mide la frecuencia de aparición de cada una de ellas. Para la formación de las clases recurren a la construcción de Clúster, basados por ejemplo en un algoritmo conocido como K- Medias, es una técnica basada en un *centroide,* el cuál es un punto central calculado por la media de los objetos o palabras asignada a un Clúster. La diferencia entre un objeto $\boldsymbol{p} \in C_i$, donde C_i es el Clúster, es una medida para la distancia $dist(\boldsymbol{p}, C_i)$; esta es la distancia Euclidiana entre dos puntos *x*, *y*. En otras palabras, se forman agrupaciones basados en que cada observación pertenece al grupo más cercano a la media.

La calidad del Clúster puede ser medida por la variación del mismo, la cual se calcula como la suma del cuadrado del error entre todos los objetos del Clúster C_i, y el centroide c_i. La ecuación es:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, C_i)^2$$

Ésta es una función objetiva que permite hacer un Clúster compacto y lo más separado posible uno de otro Clúster.

Si el número de grupos k y espacio d se fija, el problema puede ser resuelto en el tiempo $O(n^{dk+1} \log n)$, donde n es el número de datos. El algoritmo K- medias define el centroide de un grupo como el valor medio de los puntos dentro de la agrupación, este algoritmo selecciona aleatoriamente k de los datos en D, cada uno de los cuales representa un grupo o centro. Se compara con los demás objetos pertenecientes a los objetos restantes, y se le asigna a la agrupación o el Clúster más similar, basado en la distancia Euclidiana. Este proceso se repite, es decir, es iterativo; se calcula la media utilizando los objetos asignados a la agrupación, se resigna a otros grupos calculando los centros; este proceso continúa hasta que la asignación es estable, es decir, los centros recalculados son iguales a los anteriores y hay una asignación definitiva de Clústers.

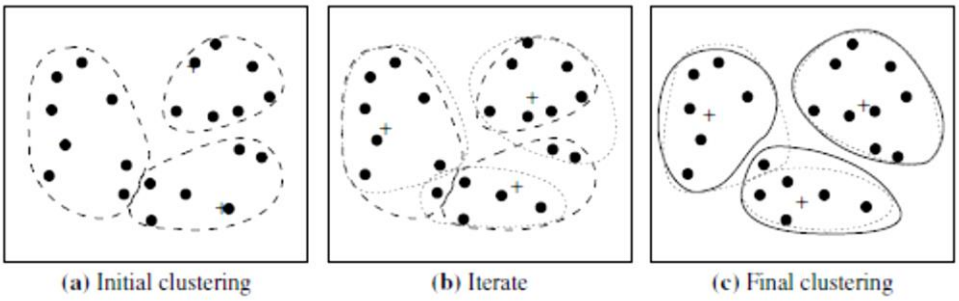


Fig. 3 Algoritmo de K – medias (Han. J., et al, 2012).

Como se muestra en la Figura 3, inicialmente se forman los Clústers, luego se calcula la media de cada uno de los Clústers comparando los objetos de los Clústers con esta, y se calculan las distancias entre éstos y las medias iterativamente, hasta que las medias sean las mismas en cada Clúster y no cambien con las iteraciones. Esto no garantiza que se converja en el óptimo global, pues siempre se termina en el óptimo local. Estos resultados dependen del Clúster inicial que se elija para que este algoritmo de buenos resultados; lo común es

correr el algoritmo múltiples veces con diferentes centros de Clúster. El método de K –medias puede ser aplicado solamente cuando la media de un conjunto es definida.

Una variación del método de k-media es el *k- medoides*, el cual emplea una medida de disimilitud con base en la frecuencia y los *medoides* de los Clúster. Este proceso no es útil si los Clúster tienen formas no convexas o el tamaño de los Clúster es muy diferente. Por otro lado, esto es susceptible al ruido y a los datos con valores atípicos (Han, J., et al, 2012).

El método de k-means para la construcción de Clúster es implementado en esta investigación para proveer el número de temas a buscar dentro de las transcripciones, la calidad del método depende de qué tanto ruido exista en el proceso, por ejemplo, siempre se usan artículos lingüísticos, como por ejemplo ‘‘el’’; estas palabras son irrelevantes dentro del documento, por lo que se deben dejar solo las palabras esenciales, eliminando artículos, palabras cortas, entre otras, que permitan limpiar el documento. Adicionalmente, se deben tener en cuenta métodos de transformación que permiten reducir las dimensiones del corpus para optimizar las relaciones entre el léxico, estos métodos se conocen como: *Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (pLSA), LDA Latent Dirichlet Allocation, y Non-negative Matrix Factorization (NMF).*

En la implementación del Clúster se debe tener en cuenta la escogencia de las características para formar los grupos, uno de estos es la *selección basada en la* ***frecuencia*** *de palabras en los documentos*, en este caso, se usa la frecuencia como forma de escoger palabras vacías o *stop words,* palabras tales como en, o, un, de, etcétera, o en algunos casos, errores de ortografía son considerados ruido, ya que permiten dejar las menos frecuentes, las cuales son relevantes para la formación de los Clúster. Para la escogencia de estos términos es importante tener un sistema de calificación de cada palabra para lograr relacionar varios documentos por las medidas de similitud, como lo son la medida coseno u otra medida que se denomina *‘Fuerza’ s(t),* que permite relacionar pares de documentos al azar de la siguiente forma:

$$s(t) = \frac{Numero\ de\ pares\ en\ las\ que\ t\ ocurre\ al\ mismo\ tiempo.}{Numero\ de\ pares\ en\ las\ que\ t\ ocurre\ la\ primera\ ves}$$

Este término, *fuerza s(t)*, se puede probar para un término el cual esté distribuido aleatoriamente en el documento con la misma frecuencia; si este término está por debajo de dos desviaciones estándar más grandes del término al azar, este se elimina.

Se requiere para este procesamiento $O(n^2)$ operaciones, lo cual lo hace poco práctico y podría presentar el aumento de dimensionalidad de los datos, este factor es decisivo para no hacer uso de este método directamente para la obtención de temas, se presenta a continuación el modelo Asignación latente de Dirichlet (*Latent Dirichlet Allocation LDA)* donde se exponen las ventajas frente a los demás modelos asociativos.

2.3. Asignación latente de Dirichlet (Latent Dirichlet Allocation LDA)

Es un modelo para extracción de Temas dentro de un conjunto de documentos, plantea que todos los documentos son una mezcla de diferentes Temas, en donde estos temas están formados por palabras con una alta probabilidad, permitiendo formar documentos más reales. Estos Temas se encuentran a través de una distribución de Dirichlet, la cual es una familia de distribuciones continuas, parametrizadas por un vector α, se conoce esta distribución como una generalización multivariada de la distribución Beta, la cual permite representar proporciones. Se denomina latente, porque es una variable que no es observable si no que esta se debe inferir a partir de los análisis.

Inicialmente se hace una asignación temporal de las palabras a cada uno de los Temas, de acuerdo con la distribución de Dirichlet, esto se conoce como modelo a priori, luego se recalcula la probabilidad de Dirichlet, teniendo en cuenta que tan frecuente es la palabra en cada documento y que tan frecuente es el Tema en el documento. Este proceso es iterativo que se detiene cuándo a analizado cada palabra en el documento, conocido esto como probabilidad a posteriori ("Latent Dirichlet Allocation", 2017).

LDA y la distribución *a priori*, que toma alguna incertidumbre antes de tomar los datos; *una a posteriori*, aplicando el Teorema de Bayes. La probabilidad a priori se multiplica por la verosimilitud, y al normalizar se obtiene la distribución de probabilidad a posteriori, la cual

es la probabilidad de la distribución condicional dados los datos. Los parámetros de las distribuciones a priori son llamados hiperparámetros, para distinguirlos de los parámetros del modelo.

Por ejemplo, si se está usando una distribución beta para modelar la distribución del parámetro ρ, entonces:

> ρ es un parámetro de una distribución Bernoulli, y α y β son parámetros de la distribución a priori (distribución beta), y por lo tanto hiperparámetros ("Probabilidad a priori", 2017).

Para analizar el modelo se deben precisar las siguientes definiciones (Blei, Y.Ng & Jordan, 2003):

- ***Palabra* w**: Es la unidad básica de datos discretos, definida como elemento de un índice de vocabulario {1,…., V}. Las palabras se representan empleando unidades bases como vectores, los cuales tienen un único componente como uno y el resto son ceros. La posición de la palabra v en el vocabulario total es representada por un vector w de tal forma que $w^v = 1$ y $w^u = 0$, para u diferente de v.
- ***Documento:*** es una secuencia de N palabras denotadas por el vector **w**=(w_1, w_2,…., w_N).
- ***Corpus:*** es una colección de M documentos denotada por M={ $\mathbf{w_1}$, $\mathbf{w_2}$,…., $\mathbf{w_M}$ }.
- ***Tema:*** z_n

El documento θ, no es directamente vinculado a la palabra **w;** estas relaciones están dadas por variables ocultas y, la variable z, se involucra para representar que en tanto de un tema en particular, se emplea una palabra en el documento. Por lo que es necesario introducir los parámetros α y β sobre el documento y la distribución de temas a buscar, ***por lo que es capaz de procesar documentos invisibles***. La estructura del modelo LDA, permite la interacción de

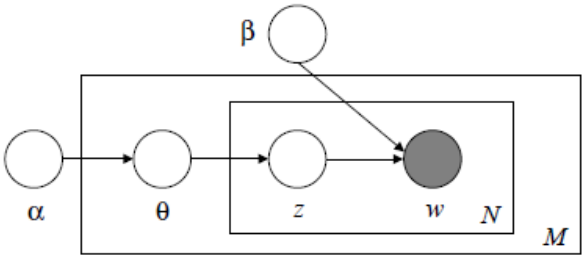


Fig. 4. Representación gráfica de LDA.

palabras observadas en el documento con distribuciones estructuradas de modelos de variables oculta (Berry & Kogan, 2010).

El modelo LDA asume el siguiente proceso generativo para cada documento **w** en un corpus, dado que el proceso Dirichlet está basado en una función de distribución donde el rango esta dado por distribuciones como Poisson, Multinomial, Dirichlet, para N que son las palabras, θ es el documento, como se muestra a continuación (Blei, Y.Ng &Jordan, 2003):

1. Escoge N ~ Poisson (ξ).
2. Escoge $\theta \sim$ Dir(α), distribución de *Dirichlet.*
3. Para las N palabras w_n: a) escoge un tema $z_n \sim$ Multinomial (θ), b) escoge una palabra w_n de $p(w_n|z_n, \beta)$, una probabilidad multinomial condicionada sobre un tema z_n

La dimensionalidad k de la distribución de *Dirichlet,* que se asume conocida y mezclada. La probabilidad de las palabras es parametrizada por una matriz β, k x V, donde β es considerada una cantidad mezclada que debe ser estimada $\beta_{ij} = p(w^j = 1|z^i = 1)$, el asumir una distribución de *Poisson,* permite tener una mayor cantidad de información. Adicionalmente, N es una variable independiente de todas las otras que eran generadas θ y **z,** estas son variables auxiliares de las cuales generalmente se ignora la aleatoriedad (Blei, Y.Ng & Jordan, 2003).

Dados los parámetros α y β, la distribución donde se unen la mezcla de temas θ, un conjunto de N temas **z,** y un conjunto de N palabras **w** está dada por:

$$p(\theta, z, w|\alpha, \beta) = (\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

Donde $p(z_n|\theta)$ es simplemente θi. Después de integrar sobre z, se llega a la probabilidad de un corpus.

$$p(\theta, z, w|\alpha, \beta) = \prod_{d=1}^{M} \int (\theta|\alpha) \left(\prod_{n=1}^{Nd} \sum_{zdn} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d$$

Los parámetros $\alpha\ y\ \beta$, parámetros del nivel del corpus, se suponen una vez muestreados en el proceso de generación del corpus. La variable θ_d, es una variable de nivel del documento, muestreado una vez por el documento. Las variables z_{dn} y w_{dn}, son variables de nivel de palabras, y son muestreadas una vez por cada palabra en cada documento de definiciones (Blei, Y.Ng & Jordan, 2003).

El conjunto infinito de variables aleatorias {z1, …, zn} se denominan *intercambiables*, *si la distribución de la probabilidad es invariante a la permutación*, por lo que LDA asume que las palabras son generadas por temas y que estos temas son infinitamente *Intercambiables* dentro del documento, dando la probabilidad de secuencia de palabras y temas de la siguiente forma (Blei, Y.Ng & Jordan, 2003):

$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n) \right) d\theta$$

Donde θ es el parámetro aleatorio multinomial sobre los temas.

2.3.1 Estimación de parámetros.

Para estimar los parámetros se empleará el *método de Bayes*, donde se buscan los parámetros α, β, los cuales maximizarán la probabilidad logarítmica de los datos. Existe un proceso variacional denominado EM, el cual maximiza los limites inferiores con respecto a los parámetros variacionales $\gamma\ y\ \phi$, conocidos como parámetros de *Dirichlet.* Así, para valores mezclados de los parámetros variacionales, se maximizan los límites inferiores con respecto al modelo de parámetros α, β. Este proceso EM se procesa de la siguiente forma (Blei, Y.Ng & Jordan, 2003):

- E – Step: Para cada documento encuentra el valor optimizado de los parámetros variacionales $\{\gamma_d^*, \phi_d^*\}$, en estos parámetros se encuentran la divergencia *Kullback* -

Leibler (KL) entre la distribución variacional y la probabilidad posterior $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$

$$\phi_{ni} \propto \beta_{iwn} \exp\{E_q[\log(\theta_i)|\gamma]\}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

La esperanza multimodal se computa de la siguiente forma:

$$E_q[\log(\theta_i)\,|\gamma] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_i\right)$$

Donde Ψ es la primera derivada de la función *log*, la cual es calculada por aproximación de Taylor.

- M – Step: debe encontrar la máxima probabilidad estimada con suficientes estadísticas para cada documento bajo una aproximación posterior, la cual está dada en el paso E. Lo anterior, se logra maximizando los resultados de los límites inferiores con respecto al modelo de parámetros α, β.

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{Nd} \phi^{*}_{dni}\, w^{i}_{dn}$$

2.3.2 Suavizamiento (*Smoothing*).

La cantidad de palabras creadas en un documento nuevo pueden causar problemas; en ocasiones, los nuevos documentos formados pueden contener palabras que no aparecen en ninguno de los documentos que conforman el corpus.

La máxima verosimilitud estimada de los parámetros multinomiales, una vez asignada la probabilidad cero para algunas palabras, y la probabilidad cero para algunos documentos, muestra que la solución a este problema es la suavización de los parámetros multinomiales, asignando una probabilidad a todo el vocabulario. El parámetro β es tomado como una matriz aleatoria k X V, donde se asume cada fila independiente de una distribución de Dirichlet intercambiable; el parámetro β será una variable aleatoria dotada de distribución

posterior condicionada por los datos, la cual está dada por la expresión (Blei, Y.Ng & Jordan, 2003):

$$q(\beta_{1:k,}z: M, \theta_{1:}M|\lambda, \phi, \gamma) = \prod_{i=1}^{k} Dir(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d\left(\theta_{d,}z_d|\phi_d, \gamma_d\right)$$

Donde $q_d\left(\theta_{d,}z_d|\phi_d, \gamma_d\right)$ es la distribución variacional definida por LDA. Se introduce un nuevo parámetro variacional λ:

$$\lambda_{i\ j} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^{*}\, w_{dn}^{i}$$

La iteración de estas ecuaciones a la convergencia produce una distribución posterior aproximada sobre $\beta, \theta\ y\ \boldsymbol{z}$, donde η es un hiperparámetro intercambiable de Dirichlet, así como el hiperparámetro α anterior. El enfoque para estos hiperparámetros será emplear Bayes de forma empírica para encontrar la máxima verosimilitud de estos parámetros (Blei, Y.Ng & Jordan, 2003).

Un ejemplo se encuentra en el libro de *minería de texto* de Berry & Kogan (2010), en el que se muestra la aplicabilidad de LDA en documentos de Wikipedia, considerada una de las bibliotecas virtuales más grandes; se escogen artículos que conserven igual semántica y, en general, artículos que de alguna forma estén relacionados entre sí. El número de documentos se establece como **ρ**. LDA forma dos distribuciones de Wikipedia, la distribución entre temas y palabras $\boldsymbol{\phi}$ y la distribución documento - tema $\boldsymbol{\theta}$,

$$\boldsymbol{\phi_{ik}} = \frac{\boldsymbol{C_{wi,k}^{W\ K}} + \boldsymbol{\beta_i}}{\sum_{v=1}^{W} \boldsymbol{C_{v,k}^{W\ K}} + \boldsymbol{\beta_v}}, \qquad \boldsymbol{\theta_{mk}} = \frac{\boldsymbol{C_{m,k}^{D\ K}} + \boldsymbol{\alpha_k}}{\sum_{j=1}^{K} \boldsymbol{C_{m,j}^{D\ K}} + \boldsymbol{\alpha_j}},$$

Donde m es el índice del artículo de Wikipedia, dentro de los artículos relacionados. $\boldsymbol{C_{wi,k}^{W\ K}}$, es el número de veces que la palabra *i* está asignada al tema *k* y $\boldsymbol{C_{m,k}^{D\ K}}$, es el número de veces que el tema *k* se asigna a alguna palabra tokenizada en un artículo *m*. Estas son consideradas distribuciones a priori, las cuales se actualizan empleando pruebas en el documento. Así, la

distribución palabra y tema $\boldsymbol{\phi}$ se actualiza a la distribución $\widehat{\boldsymbol{\phi}}$, la distribución tema – documento $\widehat{\boldsymbol{\theta}}$ se toma desde cero a partir de las pruebas al documento.

$$\widehat{\boldsymbol{\phi}}_{ik} = \frac{\boldsymbol{C}_{wi,k}^{W\,K} + \boldsymbol{C}_{-wi,k}^{W\,K} + \boldsymbol{\beta}_i}{\sum_{v=1}^{V} \boldsymbol{C}_{v,k}^{W\,K} + \boldsymbol{C}_{-vi,k}^{W\,K} + \boldsymbol{\beta}_v}, \qquad \widehat{\boldsymbol{\theta}}_{dk} = \frac{\boldsymbol{C}_{m,k}^{D\,K} + \boldsymbol{\alpha}_k}{\sum_{j=1}^{K} \boldsymbol{C}_{-d,j}^{D\,K} + \boldsymbol{\alpha}_j},$$

Donde **d,** es el índice de documento de prueba, $\boldsymbol{C}_{-vi,k}^{W\,K}$ es el número de veces que el tema k es asignado a alguna palabra en el documento de prueba d, por lo tanto, el modelo generativo está influenciado por los temas de Wikipedia.

Los modelos que subyacen al modelo LDA, son el LSI y el modelo pLSI, estos se enrutan hacia la obtención de Temas, el modelo LSI *Latent Semantic Indexing*, este es un método clasificatorio, el cual tiene categorías predefinidas, a diferencia de LDA, las cuales son en función de la similitud con el contenido de cada categoría, hace una comparación del concepto general de los textos, inclusive si se usa en motores de búsqueda arroja resultados en diversos idiomas. El método (LSI) *Latent Semantic Indexing*, matemáticamente hace un análisis de Componentes Principales para un conjunto de datos d-dimensional, CP construye la matriz de covarianza C simétrica *d x d*, donde (i , j) es la entrada de la matriz C. La matriz es semi definida positiva y puede ser diagonalizada:

$$C = P \cdot D \cdot P^T$$

Donde P es una matriz que contiene en sus columnas vectores propios *orthonormales*, C y D son matrices diagonales que contienen los correspondientes valores propios. Los vectores propios forman una base, representan una nueva base *orthonormal*, la cual permite la representación de datos. En este contexto, los valores propios corresponden a la varianza cuando los datos están sobre el eje de la base. Así, con las nuevas bases se reduce la dimensionalidad, esto sucede porque la varianza es muy pequeña en estas dimensiones. Este método de CP es comúnmente usado para medir la similitud. LSI es similar al análisis de componentes principales CP, se emplea en LSI la matriz de covarianza. Específicamente, A es *n x d* matriz de términos y de documentos, en donde (*i, j*), son las entradas de la frecuencia normalizada para el término *j* en un documento *i* (Witten & Frank, 2005).

Este modelo LSI emplea un valor de descomposición de la matriz X para identificar un subespacio lineal en el espacio de *tf – idf*, esto es la frecuencia de cada uno de los términos tf, y idf es la frecuencia inversa de documento que es la ocurrencia de cada término en la colección de documentos. Según Deerwester, en Blei, Y.Ng & Jordan, (2003), esto es una combinación lineal de frecuencias que permite capturar aspectos y nociones lingüísticas esenciales como sinónimos y polisemias (palabra con diferentes significados) (Blei, Y.Ng & Jordan, 2003), este método de pesos de frecuencias es empleado en los motores de búsqueda como elemento esencial para medir la importancia del documento encontrado dada la búsqueda del usuario (Es.wikipedia.org, 2017).

El otro modelo mejorado de LSI fue, el modelo pLSI (*probabilistic LSI*), el cual plantea que cada palabra en un documento proviene de un modelo de mezclas, donde los componentes mezclados son variables multinomiales aleatorias, las cuales pueden ser vistas como una representación de temas, y así, ***cada palabra se desprende de un tema y diferentes palabras en el documento generan diferentes temas***. Cada documento es representado como una lista de mezclas proporcionadas de componentes y se reduce a una distribución de probabilidad de un conjunto de temas (Blei, Y.Ng & Jordan, 2003).

Sin embargo, este modelo ha tenido problemas, como, por ejemplo:

- El número de parámetros en el modelo aumenta linealmente con el tamaño del corpus, lo que lleva problemas de exceso de información.
- No es claro cómo se asigna la probabilidad al documento fuera del grupo de datos.

La probabilidad de que un documento *d* y una palabra w_n, sean condicionalmente independientes dado un tema inobservable z (Blei, Y.Ng & Jordan, 2003):

$$p(d, wn) = p(d) \sum_{z} p(wn\,|z) p(z|d)$$

Esto muestra la posibilidad de que el documento contenga múltiples temas *p(z|d)*, da una mezcla de pesos de los temas para un documento en particular *d*. El valor *d* es un valor que puede tomar diversos valores aleatorios -es multinomial-, a partir de los cuales se forman los

documentos, lo cual genera un problema, ya que solo calcula las mezclas de los documentos que dependen de *d* únicamente, por lo que lo hace un modelo no generativo de documentos, ***lo cual no permite asignar probabilidades a los documentos ocultos***. Además, los parámetros de k – temas, de tamaño V y mezcla M, sobre k temas ocultos, está dada por Kv + kM parámetros, aumentando de forma lineal con las mezclas M, por lo tanto ***el modelo es propenso a sobrealimentación,*** los cuales se consideran como fallas del modelo, por lo que el modelo LDA (*Latent Dirichlet Allocation*) supera ambos problemas para tratar mezclas de temas y sus pesos probabilísticos y el no aumento de parámetros (Blei, Y.Ng & Jordan, 2003).

Un primero modelo de análisis probabilístico latente fue Probabilistic Latent Semantic Analysis (probabilidad de análisis semántico latente (pLSA) según Berry & Kogan, 2010, muestra que Thomas Hofman en 1999, introduce una aplicación de los métodos Bayesianos para modelar los documentos, la cual asocia un aspecto o clase inobservable, asociada a una variable zk, en donde cada documento d, representa cada distribución sobre cada palabra, $p(w|z)$ *, este modelo es parametrizado por la distribución de un documento d y una palabra* w_{di}

$$p(d, w_{di}) = p(d) \sum_{z=1}^{K} p(w_{di}|z)\, p(z|d)$$

Este modelo considera aspectos ocultos, asume que los documentos y las palabras a estudiar son independientes, permite asociar un peso probabilístico $p(w|z)$ *a los documentos con cada uno de los temas, su representación gráfica muestra la probabilidad (Berry, & Kogan, 2010).*

El proceso pLSA : Esboza un documento con probabilidad $p(d)$.Realiza para cada palabra *i* en un documento *d*:

a. Dibuja un aspecto latente (oculto) z_i con probabilidad $p(w_{di}|z_i)$.
b. Dibuja una palabra w_{di} con probabilidad $p(w_{di}|z_i)$.

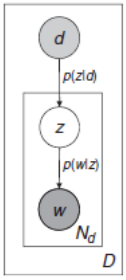


Fig. 5. Representación gráfica del proceso pLSA Berry & Kogan, 2010

Un problema con el modelo fig.5 es que principalmente no es de carácter generativo, es decir, no logra crear datos observables, donde la variable d es una variable *dummy* aleatoria, la cual genera un conjunto de documentos, y por lo tanto

surge otro problema relativo a la producción de ***una sobrecarga de formación de datos que afecta la inferencia de aspectos del modelo en documentos ocultos***. A pesar de esto, este modelo hace una gran contribución a la Minería de Texto, toda vez que estos ***problemas de aumento de dimensionalidad y estudio de variables ocultas*** dio paso a LDA. (Berry & Kogan, 2010).

Los problemas mencionados en estos modelos como el de Cluster, LSI, pLSI, *muestran que al no ser estos modelos generativos, presentan problemas con el aumento de la dimensionalidad y de los parámetros, así mismo, estos modelos no posibilitan asignar probabilidades a esas relaciones ocultas entre los textos, por lo que al ser LDA un modelo generativo, este crea hiper parámetros para lo oculto, sin causar aumento en la dimensionalidad*, además de ser un algoritmo pensado para la creación de grupos de palabras, a partir de los temas o tópicos, y construir un conjunto de temas a partir de los documentos. Por lo que este modelo probabilístico es el implementado en la extracción de Temas dentro de los manuscritos.

2.4. Herramientas computacionales.

En este trabajo se busca una herramienta computacional que sea amigable y además este en constante renovación en métodos de extracción de información por medio de la minería textual.

2.4.1. KNIME.

Knime es una plataforma de código abierto como herramienta de colaboración e investigación. Debido a que este producto tiene que procesar e integrar grandes cantidades de datos diversos, los desarrolladores se adhirieron a rigurosos estándares de ingeniería de software para crear una plataforma sólida, modular y altamente escalable que abarca varios cargamentos de datos, transformación, análisis y modelos de exploración visual. está escrito en Java y está basado en Eclipse, el entorno de desarrollo de software multilingüe de fuente abierta que comprende un entorno de desarrollo integrado (IDE) y un sistema extensible de plug-in (KNIME Open Source Story | KNIME, 2017).

Adicionalmente esta herramienta es bastante fácil de implementar, su estructura permite que se arme un flujo en una página en blanco, simplemente arrastrando los nodos que ya están predefinidos y se conecten con flechas de entrada y salida, lo único que el usuario debe hacer es programar cada uno de los nodos con los parámetros requeridos, por lo que esto hace que cualquier persona con mínimos conocimientos computacionales acceda a este lenguaje, también cuenta con ejemplos prácticos de flujos dependiendo el análisis que se quiera hacer, como se muestra en la fig 4, una lista de estos. Si algún historiador quiere hacer uso de esta herramienta se encontrará con un lenguaje amigable, sin necesidad de ser experto en lenguaje de programación.

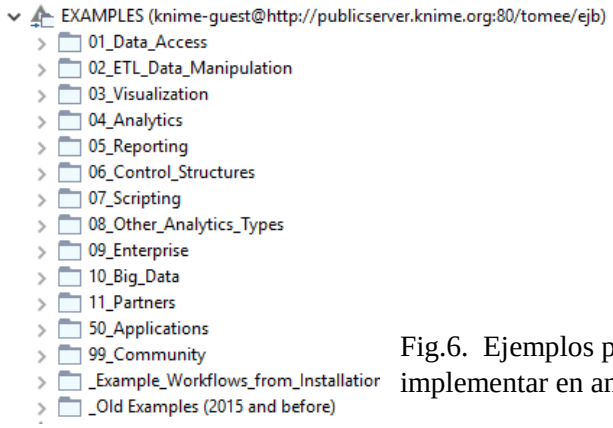


Fig.6. Ejemplos proporcionados por KNIME de flujos a implementar en análisis de datos.

El proceso de *Minería de Texto* según Knime, 2013, debe llevarse a cabo de la siguiente manera como se muestra en la fig. 7:

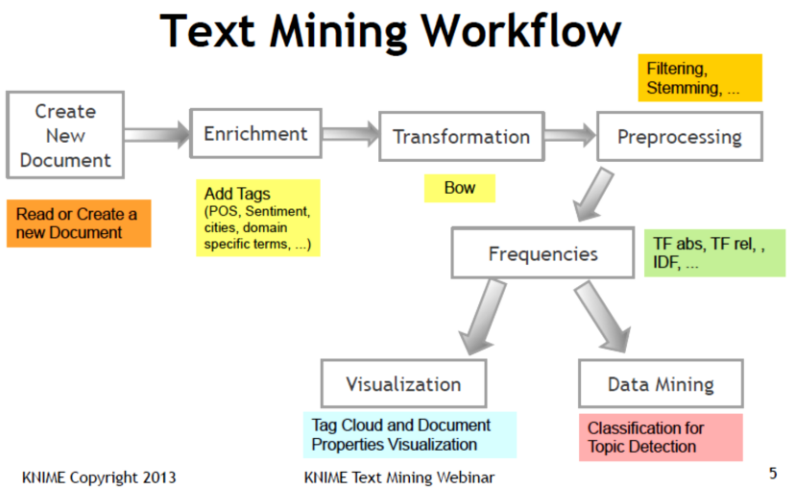


Fig.7. Flujo de trabajo Minería de texto (KNIME, 2013)

A continuación, se explica cada una de las partes del flujo de trabajo existentes en la fig. 5.

- ✓ **Textos:**

  Crear un documento en este caso el Corpus – conjunto de textos o datos – elegido para la investigación son las transcripciones realizadas por los estudiantes de Coursera, el cual se encuentra en formato Xls.

- ✓ **Enrichment (enriquecimiento) :** en esta fase se debe Etiquetar cada una de las palabras, a partir de diccionarios semánticos que permiten asignar a cada palabra si es un adjetivo, nombre, verbo, pronombre un artículo, etcétera, generalmente los Software tienen predeterminados diccionarios con los que se compara y etiqueta la

información en idiomas como : ingles, francés, alemán, español, sin embargo, para esta investigación dichos diccionarios eran insuficientes por estar en *español actual*, y recordemos las *actas capitulares* se encuentran escritas en *Castellano antiguo*, por lo que se hace necesario crear diccionarios propios de nombres, lugares, adjetivos y verbos. La obtención de estos diccionarios fue uno de los objetivos principales de la investigación de Socha, Martinez & Medina (enero 2017), implementaron un modelo de *extracción en Phyton y el software KNIME* como se muestra en el código del trabajo que desarrollaron y se encuentra en el ANEXO de este trabajo, donde obtuvieron los diccionarios propios de *las actas capitulares de Plasencia.*

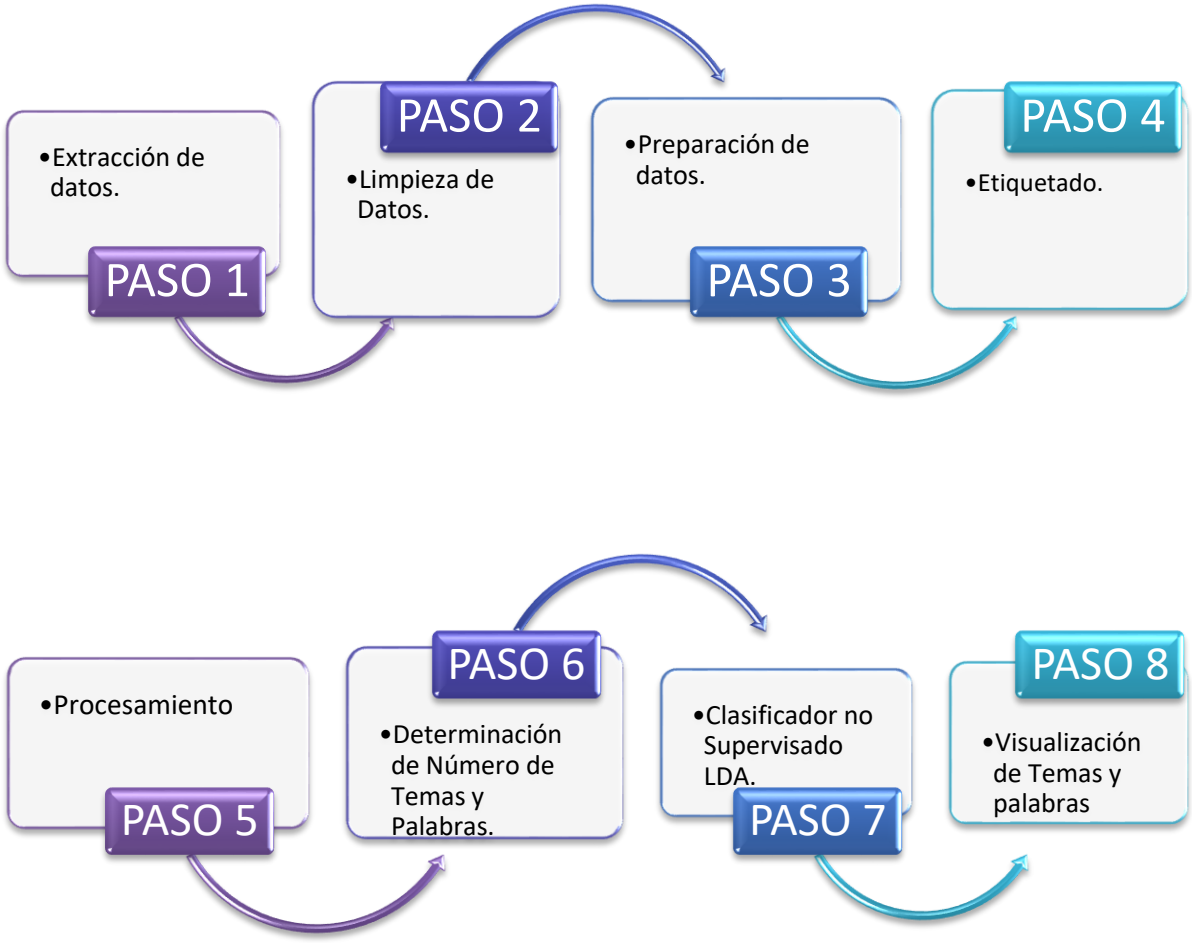
- ✓ **Transformation (Transformación):** una vez *etiquetadas* las palabras del *corpus*, se crea una *bolsa de palabras - Bow (bag of words)* - , las cuales son las que se deben someter al análisis de MT minería textual.
- ✓ **Preprocessing (procesamiento) y frecuencia:** en este paso se procede a quitar todas aquellas palabras que no aportan al análisis de los documentos, conectores, artículos, palabras repetidas, lo que se denominan *Stop Words o palabras vacías*, en este filtrado se calcula la frecuencia de aparición de cada uno de los términos, para así descartar por número de *frecuencia palabras*, por ejemplo si aparecen 1000 veces el articulo el, se descarta de inmediato, o se predeterminan que palabras con cierto número de caracteres sean removidas, ya que por lo general son palabras vacías como por ejemplo: ¨una¨, ¨las¨, etc.

3.METODOLOGÍA E IMPLEMENTACIÓN.

A continuación, se presenta un esquema de la metodología que sigue el presente trabajo investigativo.

3.1.Detección de Temas (topics) implementando Minería Textual.

Se muestran las etapas que permitieron llegar a los temas dentro de las transcripciones.



El esquema muestra, la metodología implementada para llegar a la visualización de los temas dentro de los manuscritos, Paso 1, se realizó la extracción de las transcripciones dentro de la base de datos de la investigación del Dr. Martínez. Paso 2, se hace la limpieza respectiva de

información innecesaria dentro de las transcripciones (nombre de los estudiantes, números de identificación, códigos internos). Paso 3, preparación del Corpus del documento en Excel, borra las transcripciones repetidas, se quitan las columnas no necesarias. Paso 4, se pasa por un etiquetador, Paso 5, en el procesamiento se dejan solo términos importantes, para luego pasar a determinar cuántos temas serán a clasificaren el Paso6. En el Paso 7 se pasa por el clasificador no supervisado LDA, para finalmente obtener los temas con la respectiva asignación de palabras, según la probabilidad.

3.2. Descripción de cada Paso.

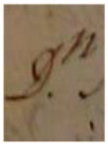
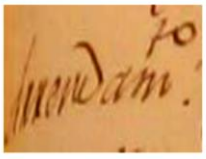
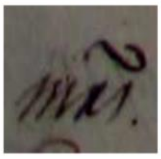
Se detallarán a continuación cada uno de los pasos.

3.2.1. Extracción de datos.

Como se mencionó al inicio del texto la investigación del Dr. Martínez, lo lleva a obtener el primer volumen de las Actas Capitulares, de la transcripción realizada en el siglo XIX, de las actas originales del siglo XV, con este material el equipo decide integrar a estudiantes de Paleografía, a través de cursos masivos por internet MOOCs (*Massive Open Online Course*).

Durante 12 semanas los estudiantes acceden al curso mediante la plataforma virtual de COURSERA, donde aprendían historia medieval de España y en particular de Plasencia, en donde a través de la práctica de paleografía -estudio y elaboración de transcripciones- se daban a la tarea de transcribir algunas abreviaturas, así como fragmentos completos, que aparecían en los folios. En la figura 8, se aprecian algunas abreviaturas empleadas, en

Fig. 8. Abreviaturas de algunos términos y asignación de letras del abecedario según caligrafía

Abbreviations, Accent Marks, and Special Forms of Words	
	"Don" Note: a title of honor to pay respect to people when naming them; the feminine version is "Doña".
	"Arrendamiento" Note: any word that ends with "to" in superscript ends with "-iento", a common suffix.
	"maravedis" Note: this is not a currency, but rather a unit for counting money to account for its real value for accounting purposes, independently of the currency actually used in a transaction.

español antiguo vistas en las actas. En total se lograron transcribir 600 páginas de las Actas, estas se escogen según sus calificaciones de pruebas evaluativas del estudiante en la MOOC y desempeño del curso, dando así un volumen importante de datos digitalizados, al aprender a transcribir textos manuscritos en antiguo español a caracteres conocidos, abre la posibilidad de entender lo que en estos documentos guardan.

Algunas de las recomendaciones que se hacen a los estudiantes del MOOC al hacer las transcripciones son:

- Escanear el documento de forma inicial para ver con qué se enfrenta.
- Identificar de forma inmediata las abreviaturas y números.
- Ubicar palabras repetidas, entender como son los conectores de palabras, reconocer letras y números del alfabeto.
- Reconocer las abreviaturas existentes.
- Identificar cognados (aquellos términos con un mismo origen etimológico, pero con distinta evolución fonética)
- Digitar la transcripción le permite reconocer dónde hacen falta términos.

En la figura 9, se muestra uno de los fragmentos que se le entregó a los estudiantes a transcribir, se oculta el número de folio por derechos de autor de la catedral. En 2014 se abrió un curso y a cada estudiante se le asigno algunos fragmentos de los manuscritos, para lograr obtener las transcripciones de la totalidad de los documentos, alrededor de 2130 alumnos hacen el curso y realizan la transcripción.

.

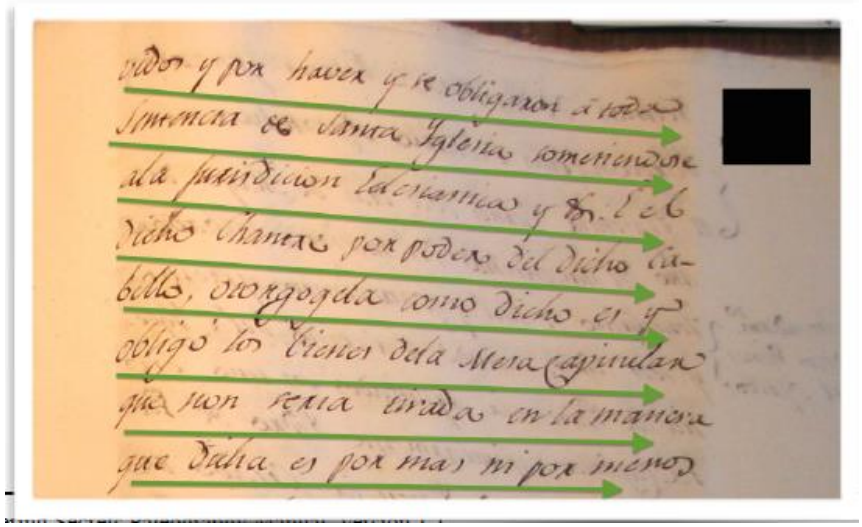


Fig. 9. Ejemplo de los fragmentos entregados a los estudiantes de Coursera.

Los resultados del curso se guardan en un archivo Excel y HTML, muestran la identificación del alumno, la transcripción de los fragmentos y la respuesta a algunas preguntas puntuales realizadas.

Son estos archivos a los cuales el Dr. Perez da acceso como conocedor y colaborador del proyecto del Dr. Roger. Existen carpetas principales, donde reposan los resultados de cada uno de los alumnos, estas carpetas muestran el resultado de una prueba escrita que se realiza; las instrucciones de las transcripciones deben seguir los siguientes pasos:

a. Primero se pide transcribir el fragmento cuya extensión debe comenzar con el nombre del grupo de estudiantes y el código de la imagen del manuscrito. Ejemplo: *Group 7 Manuscript Image B3M30.*
b. Se pide al estudiante: Si usted tiene alguna fluidez con el idioma español, también puede completar las siguientes preguntas opcionales de crédito extra. Crédito adicional Pregunta # 1: "*Si hay alguna fecha de "año" reportada en la selección, por favor registre esas. Separe todas las fechas con una coma* ","Ejemplo, 1406".
c. Si existe alguna cantidad financiera indíquela. Ejemplo, 300 Maravedies.
d. Indique nombres de lugares, si aparecen. Ejemplo: Posada de las colmenas.
e. Registre nombres de personas si allí aparecen. Ejemplo, Diego Pérez de Granada.

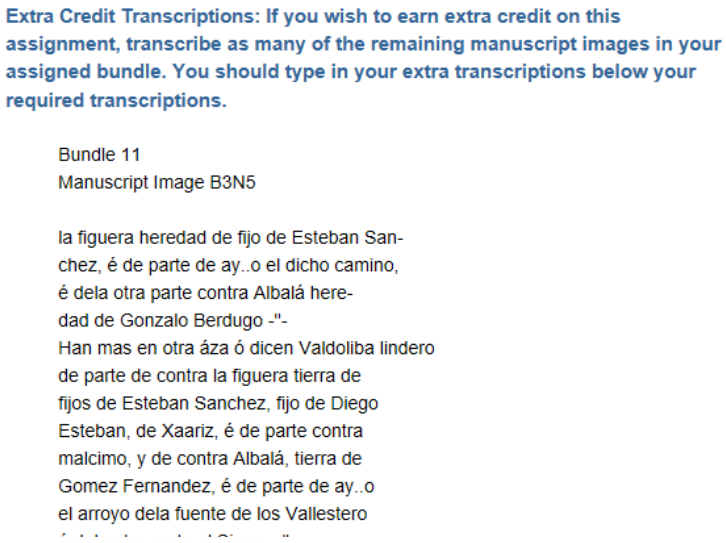


Figura 10. Ejemplo: forma de presentación de datos

Estos fragmentos son un ejemplo como se ve en la figura 10, de la presentación de las transcripciones por parte de los estudiantes, las respuestas a cada una de estas preguntas formuladas en el curso se archivaron en diversas carpetas, en las que aparecen los fragmentos y las observaciones realizadas por los pares evaluadores, los cuales dan sus observaciones, así como un puntaje para cada una de estas. A cada uno de los estudiantes se le asignó una calificación según la calidad de la transcripción, la cual tenía un puntaje máximo de 30 puntos, más algunos puntos extras de la transcripción, como decir nombres, cantidades monetarias, años, dando un puntaje máximo de 36 puntos, lo que permitió clasificar las transcripciones.

Los manuscritos están plasmados a dos caras, lo que se conoce como el ***recto y el verso,*** como se muestra en la figura 8. El recto es la parte frontal y el verso es la parte posterior de la hoja, estos términos son empleados en los campos de la codicología, paleografía, diplomática y filología (Recto y verso, 2016)

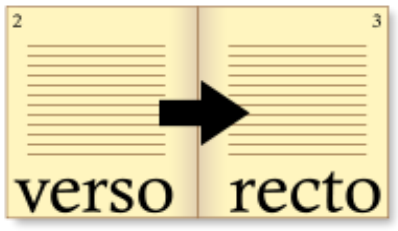


Fig 11. Ejemplo de recto y verso tomado de Wikipedia con la búsqueda recto y verso.

Adicionalmente, *cada recto y verso* los investigadores del Dr. Martínez, *segmentaron como arriba y abajo cada página*, por ejemplo, del verso 1 hay arriba y abajo y del recto de igual forma, por lo que la tarea adicional fue la reconstrucción de cada una de las páginas (recto y verso) de los manuscritos.

El equipo de trabajo de investigación del Dr Martínez, entrega la siguiente organización de los fragmentos de la transcripción, están dadas por el número de FOLIO, el cual fue un número que se le asignó a cada una de las páginas de las actas, seguido de la letra *r o v*, indicando si es recto o verso y por último indica si es el trozo de arriba o abajo.

Manuscript Image #	Libro de Actas (Folio #/recto or verso)	Section of Folio (Top/Arriba or Bottom/Abajo)
"B1C1" a "B1C100"	"Folio 300r" a "Folio 399r"	Bottom/Abajo
"B1D1" a "B1D100"	"Folio 299v" a "Folio 399v"	Bottom/Abajo
"B1E1" a "B1E100"	"Folio 300r" a "Folio 399r"	Top/Arriba
"B1F1" a "B1F100"	"Folio 299v" a "Folio 399v"	Top/Arriba
"B2R1" a "B2R101"	"Folio 201r" a "Folio 299r"	Top/Arriba
"B2S1" a "B2S102"	"Folio 200v" a "Folio 298v"	Bottom/Abajo
"B2T1" a "B2T102"	"Folio 200v" a "Folio 298v"	Top/Arriba
"B2U1" a "B2U101"	"Folio 201r" a "Folio 299r"	Bottom/Abajo

Como se muestra los folios del 300 al 399, **r** significa el **recto** de la página, la letra C indica que son todos los segmentos de la parte de abajo de la hoja, mientras que los que están marcados con la letra E, del mismo folio, son las partes de arriba. Por lo que estos fragmentos constituyen 100 páginas, de esta forma la asignación de los códigos va como sigue:

FOLIO	**CÓDIGO**
0-99	B4
100 – 199	B3
200 – 299	B2
300 – 399	B1
400 – 499	B5

Por lo que en su totalidad debe hacer la reconstrucción de 500 páginas de las actas capitulares.

Esta parte de la investigación fue una de las más complejas, dado que era un trabajo de extracción de información en la cual se debía por métodos computacionales, para conocedores en el tema. Simultáneamente se realiza un trabajo enfocado a la extracción de textos en la Escuela de Ingenieros de Bogotá, se desarrolla un software para hacer una extracción más meticulosa de las transcripciones de los estudiantes participantes en Coursera, dirigida por el Dr. Ignacio Pérez titulada *¨Minería de texto histórica - colaboración al proyecto: Revealing Cooperation and Conflict Project¨*, (Socha. D, Martinez. J, Medina. C, enero de 2017), *en este trabajo se busca implementar métodos computacionales para la extracción de datos cualitativos como: nombres, lugares, terminología de la época, empleando el software Python 3, con el cual generaron una rutina para la extracción y*

limpieza de datos – Anexo- en donde se logra la extracción de todas las transcripciones, manteniendo la sintaxis original de las mismas.

Por lo que, del trabajo de Socha, Martinez & Medina (enero 2017), se aprovechan las transcripciones de 10125 fragmentos, de todos los estudiantes participantes del curso, cabe aclarar, que su volumen es tan grande, porque de un mismo fragmento se encuentran varias transcripciones de diferentes estudiantes, cada una de estas está ligada al código B1, B2, B3, B4 y su respectiva letra, dependiendo de la ubicación dentro de los manuscritos.

Cabe notar que, en el trabajo de Socha, Martinez & Medina (enero 2017), con la implementación de este *software Python y Knime,* logran la extracción de lugares, términos, nombres de personas de la época, las cuales dan como aporte de su trabajo constituyendo así diccionarios propios de las actas procesadas, con cada una de estas características (Véase: Socha, Martinez, & Medina, 2017). Por lo que este avance primordial dentro del procesamiento automático de la información que se aplica en este trabajo, la obtención de diccionarios permitirá un mejor etiquetado de las palabras.

3.2.2. Limpieza de Datos.

Al analizar los datos obtenidos del método anteriormente mencionado se encuentra con la dificultad, que existen algunos restos de información no relevante para la investigación como la repetición de la palabra *Group*, con la que iniciaban los estudiantes las transcripciones, palabras como *Image, Manuscript, Line*, y algunos códigos del curso, aparecen de forma constante en estos datos, por no ser propias de los documentos estas causaran ruido en el análisis, por lo que fue necesario hacer un filtro desde Excel, antes del procesamiento, se quitaron alrededor de 20.000 palabras, que fueron detectadas en la pruebas previas.

3.2.3. Preparación de datos.

Esta limpieza se guardó en una hoja de cálculo en donde se tienen tres columnas una columna guarda el código del manuscrito B1, B2, B3, B4, B5, estos códigos son la partición de los fragmentos de las actas, por ejemplo, el código B5 hace referencia a las 100 primeras paginas

de las actas capitulares, el B4 a las 100 siguientes y así sucesivamente hasta terminar con B1, al lado de este código que además va ligado a una letra y un número, ej. B1C, se encuentra el resultado de la transcripción realizada por cada uno de los estudiantes. Se remueven números de identificación de los estudiantes, así como calificaciones, entre otros datos irrelevantes para la investigación.

3.2.4. Etiquetado.

Cada una de las filas se considera un documento, que constituyen en conjunto el Corpus de las transcripciones, estos se convierten en documentos, adicionalmente se procesan para identificar las palabras como verbos, nombres pronombres, etcétera, en esta parte se implementa el Procesamiento del Lenguaje Natural de la Universidad de Stanford PLN, n Tagger Part-Of-Speech (POS Tagger) es una pieza de software que lee texto en un idioma definido y asigna partes de la oración a cada palabra como lo son: sustantivo, verbo, adjetivo, etc., aunque generalmente computacional las aplicaciones usan etiquetas POS más complejas como 'sustantivo-plural', (The Stanford Natural Language Processing Group, 2003).

3.2.5. Procesamiento.

En este paso se debe remover, todo aquello que no es útil ni relevante dentro de los datos, se debe quitar la puntuación y aquellas palabras que no son aportan nada (Stop Words) como los artículos, y las abreviaciones como, por ejemplo, Don, aun, etcétera, es en este punto dónde cada nodo se debe incluir el diccionario de palabras en español antiguo, para poder hacer un filtro adecuado de las palabras a ser relevantes dentro de la investigación, dichos diccionarios se toman del trabajo de Socha, Martinez & Medina (enero 2017), para identificar las palabras vacías o ¨Stop words¨ propias del español antiguo, que se deben eliminar de las transcripciones pronombres, artículos, abreviaturas, conectores entre otros, así el documento ya está listo para crear una bolsa de palabras que forma la Tupla de información la cual se analiza, nuevamente es pasada por el POS Tagger, para etiquetar el Corpus a ser clasificado, además se calculan las frecuencias de cada termino, para hacer un

último filtro basado en la alta aparición de términos, por el método de Zifp y IDF, descrito anteriormente .

En este punto la bolsa de palabras creada está lista para ser clasificada, empleando el modelo de clasificador no supervisado, que se encuentra como un nodo en KNIME.

3.2.6. Clasificador no Supervisado, Latent Dirichlet Allocation LDA.

Esta es la ¨clasificación¨ de las actas, según la técnica Dirichlet explicada anteriormente, en este método de extracción de Temas, se procesa como cada palabra tiene una probabilidad que está relacionada con cada uno de los documentos que conforman el Corpus, el conjunto de palabras que se relacionan entre estas conforman el Tema o Topic, que arrojará la implementación del clasificador.

La configuración del nodo LDA a ser implementado como se muestra en la figura 12, muestra que previamente se debe tener claridad del número de Temas o Topics a encontrar dentro del documento, así como la cantidad de palabras que se asignaran a un Tema específico (sec.3.2.7), los parámetros para la distribución de Dirichlet α y β (sec.3.2.8), el número de iteraciones a realizar y por último el número de segmentos en que se llevarán a cabo las estadísticas de la clasificación su valor máximo es 4 (No Threads).

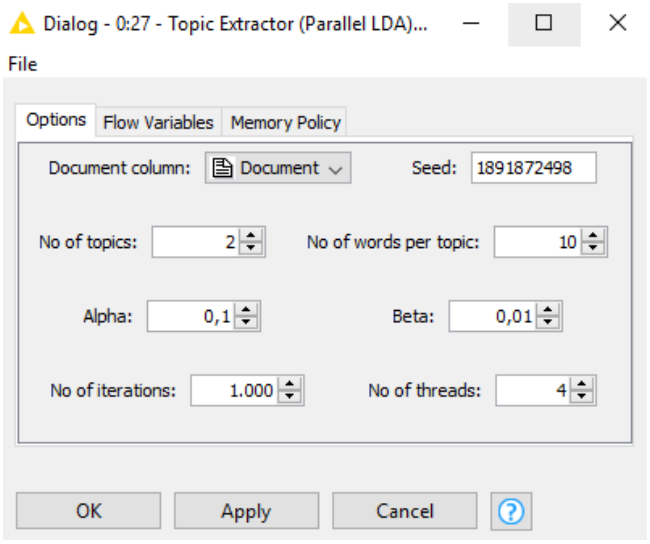


Fig.12. Cuadro de diálogo Knime para LDA

A continuación, se detalla cómo se llegó al número de temas y la asignación de los parámetros, para la implementación del modelo.

3.2.7. Determinación de Número de Temas y Palabras.

La determinación del número de temas o topics, a implementar tiene diversas formas de establecerse, estas son: análisis de la perplejidad, esta es una medida canónica de que tan bueno es el modelo, es el equivalente algebraico inverso de la media geométrica de la verisimilitud por palabra, como muestra Blei, D., Y.Ng, A. & Jordan, M, (2003), en su artículo. Otro método es buscar la media armónica de una matriz de frecuencias inversas tf – idf, se traza un gráfico de esta media armónica versus posibles cantidades de temas y el momento en que la curva desciende en ese punto se considera el óptimo de temas a ser implementados, como se muestra en el trabajo de Alvarez Ramos, (2017) .

El método ELBOW o método del Codo, según Gove, R. (2017), es la asignación de un k-means a un cluster en el conjunto de datos para un rango de valores de k de uno a diez y para cada valor de k se calcula la *Suma de Errores Cuadrados* (SSE), luego de esto se traza un gráfico de líneas de la SSE para cada valor de k. La suma de errores cuadrados, se grafican tienen en función de la cantidad de clúster, cuando la disminución se vuelve marginal, este es el número adecuado de temas a analizar. Si el gráfico de líneas se ve como un brazo, entonces el "codo", se vea el quiebre del gráfico se determina es el óptimo de temas a ser implementados.

Este método es el que se implementó en la investigación. En la figura 13 se muestra el flujo para obtener el número de tópicos o Clúster, como se mencionó anteriormente, se asigna un valor de k entre 1 y 10, en este caso de forma aleatoria se toma el 2, y se calcular la suma de los errores al cuadrado para luego pasar estos resultados a una tabla y finaliza, construyendo una gráfica en Elbow . (Gove, R. 2017)

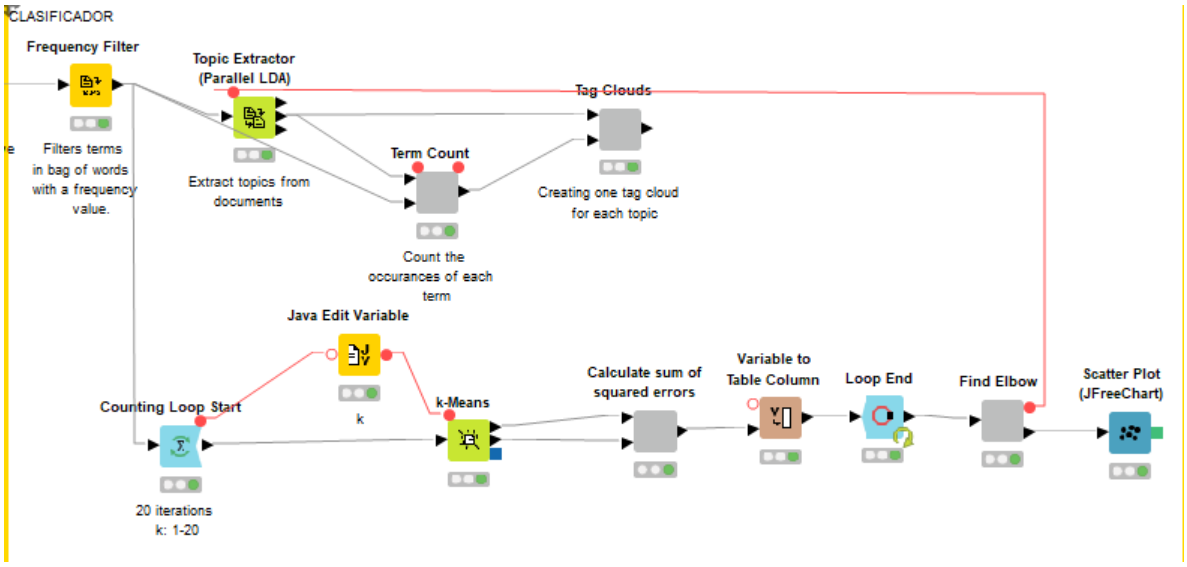


Fig.13. Flujo de Clasificación en Knime

Como se observa en la figura. 14 la suma de los cuadrados va disminuyendo a medida que aumentan el número de Topics, el punto donde se presenta este descenso muestra una cercanía con el mínimo error ya que tiende a cero, recordemos que el k -means es un valor desde el centro del Cluster, si su diferencia con la suma de los cuadrados tiende a cero, quiere decir no hay error entre el centro y k (Gove, R, 2017), esta gráfica obtenida con un conjunto de transcripciones muestra que para esos documentos el **óptimo de temas es 2,** tanto gráficamente como en un cuadro de dialogo, como se muestra en la figura. 14.

Index	Owner ID	Name	Value
0	0:764	Number of Clusters	2
0		knime.workspace	C:\Users\CATALINA\Desktop\TESIS\FLUJOS KNIME

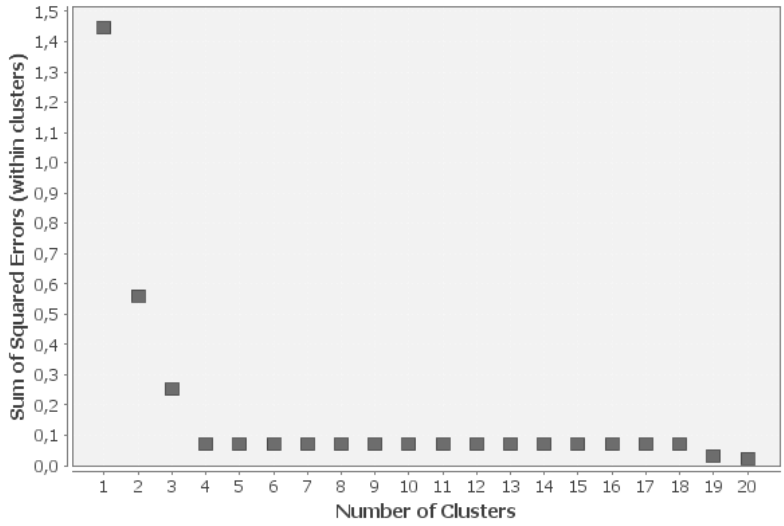


Fig.14. Resultados arrojados de proceso Ebowl

Este método se empleó a lo largo de la investigación como se mostrará en los resultados.

3.2.8. Parámetros α y β.

La elección de los parámetros se dejó como el cuadro de dialogo la arroja ya que, como menciona Griffiths & Steyvers, 2004, en su artículo de investigación, al dar un valor de β pequeño se reduce el impacto de la dispersión en los datos, lo cual afecta la granularidad estadística del modelo (tamaño de las divisiones), estas escalas de división se establecerán por este parámetro β. Es por esto por lo que entre más pequeño sea el valor de β es más posible encontrar un mayor número de Temas.

3.2.9. Visualización de Temas y palabras, Aporte de la investigación.

Finalmente se presentan los resultados de cada uno de los Temas o Topics del Corpus del documento, estas son tablas en dónde los temas están dados como TEMA 0, TEMA 1, TEMA 2..., TEMA n, etcétera, con su correspondiente conjunto de palabras asignadas, las cuales están probabilísticamente ligadas al tema según la distribución multivariada de Dirichlet, al lado derecho en los resultados arrojados siempre se encontrará el peso probabilístico de la palabra dentro del Tema, clasificado.

3.3. Flujo en KNIME.

En el trabajo de Socha, Martinez & Medina (enero 2017), proponen un flujo en KNIME para el procesamiento de los datos, se tomará como base, para de allí partir con el método Clasificador Supervisado LDA, cabe resaltar que a este se debió perfeccionar para hacer una limpieza de datos óptima. En la parte de inferior se incluye la parte de Clasificación.

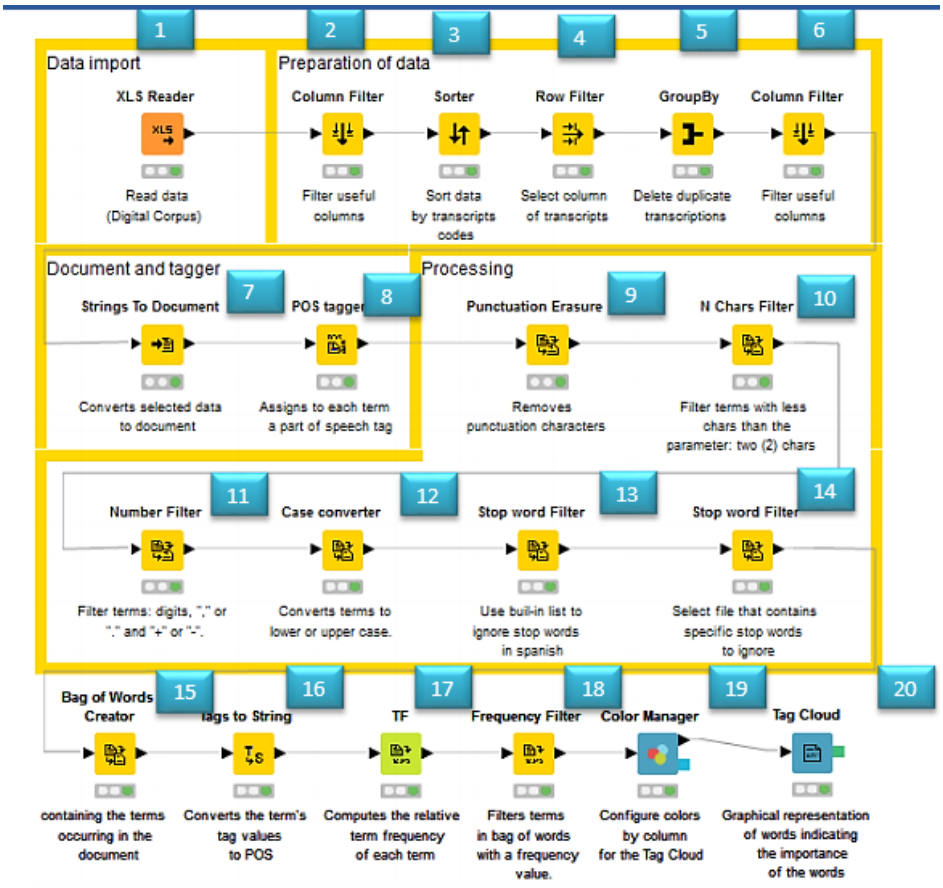


Fig. 15. Extracción de datos Flujo KNIME, investigación Minería de texto histórica - colaboración al proyecto "Revealing Cooperation and Conflict Project". de Socha Díaz, D., Martínez Serna, J., & Medina Mosquera, C. (2017)

Algunos de los nodos que ofrece Knime para la extracción de Temas y los cuales fueron implementados en la investigación se pueden ver en los anexos.

Parte clasificatoria para las actas capitulares de la catedral de Plasencia.

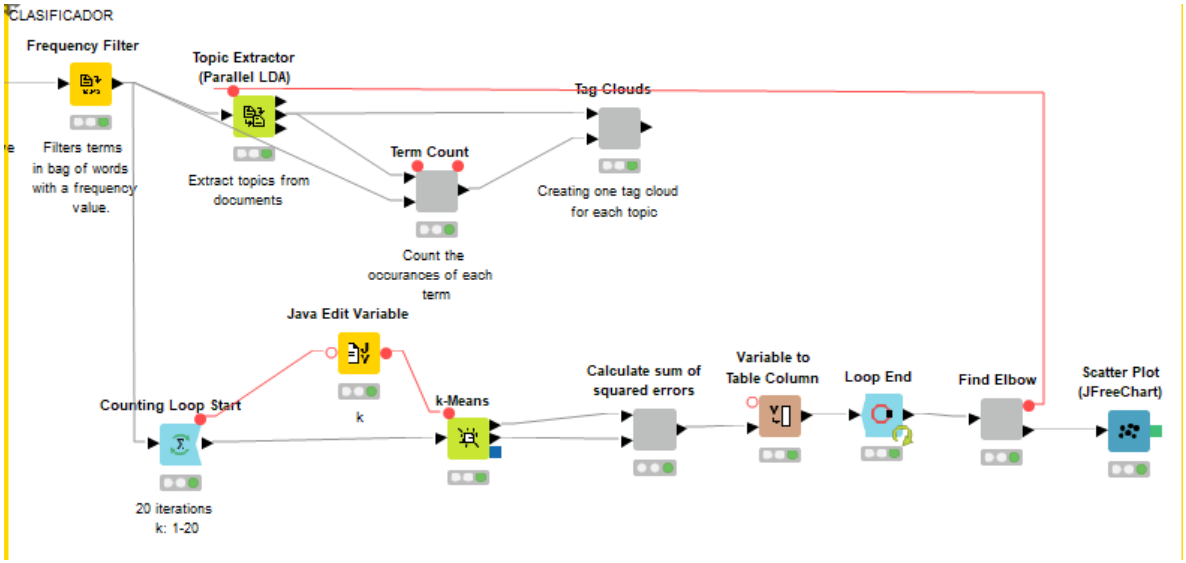
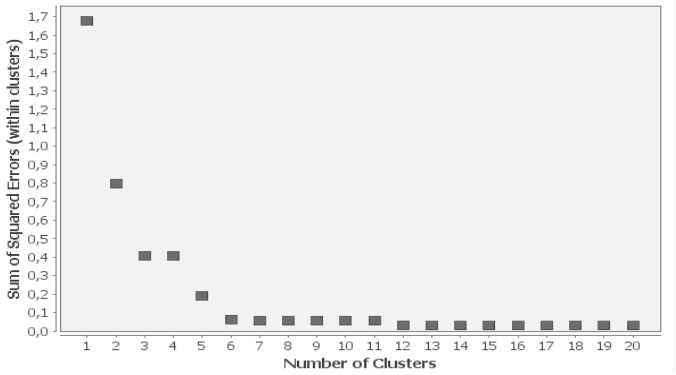


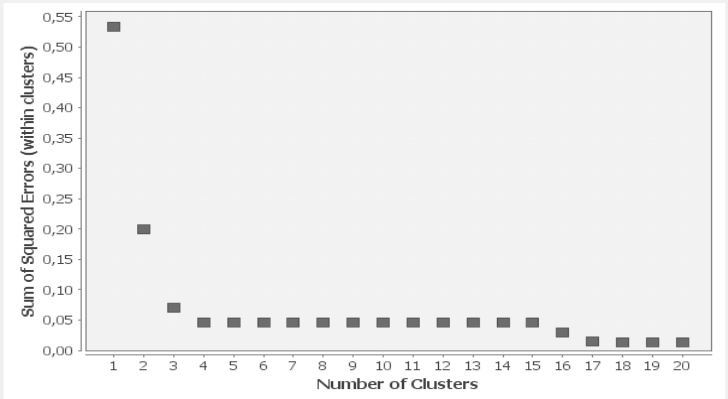
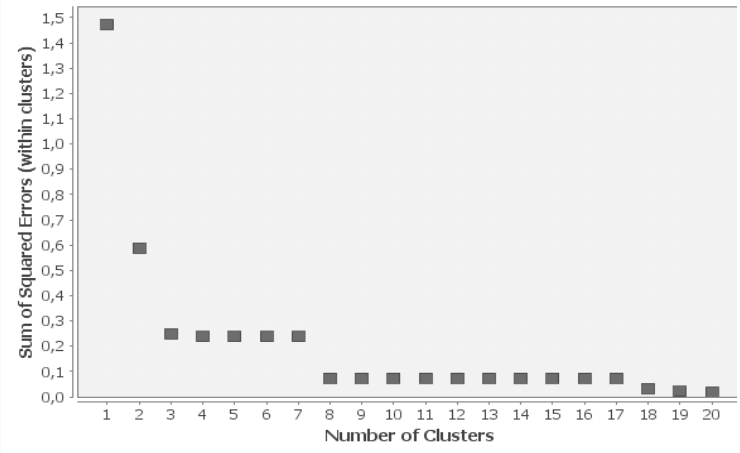
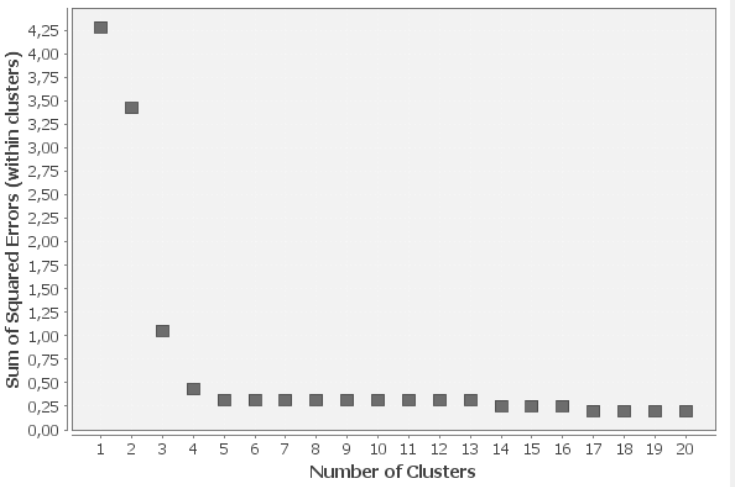
Fig.16. clasificador con LDA en Knime

4. PRUEBAS Y RESULTADOS CON LATENT DIRICHLET ALLOCATION (LDA).

4.1.Aplicación del Método ELBOW

Se hará una segmentación de las transcripciones a partir del código asignado a las actas de forma original que hicieron los investigadores del proyecto de Plasencia virtual, dividir arbitrariamente en B1, B2, B3, B4, B5, estos códigos representan división por número de páginas, como se mostró anteriormente, se debe tener presente que el orden de los manuscritos comienza desde B5 hasta B1, esto arroja desde la página 1 a la 500. Cada una de estas divisiones se estudia con el flujo creado en KNIME, después de ser estos fragmentos pre procesados como se mostró en la metodología, se hace el cálculo de Temas óptimos a ser clasificados, para después pasar al nodo LDA, arrojando la clasificación, por el método Elbow lo siguiente:

	Numero de Tópicos según ELBOW	Gráficos Suma de errores cuadrados versus número de Clúster o Topics SSE vs Topics
B5	**2**	

B4	**2**	 Sum of Squared Errors (within clusters) 0,55 0,50 0,45 0,40 0,35 0,30 0,25 0,20 0,15 0,10 0,05 0,00 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Number of Clusters
B3	**2**	 Sum of Squared Errors (within clusters) 1,5 1,4 1,3 1,2 1,1 1,0 0,9 0,8 0,7 0,6 0,5 0,4 0,3 0,2 0,1 0,0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Number of Clusters
B2	**3**	 Sum of Squared Errors (within clusters) 4,25 4,00 3,75 3,50 3,25 3,00 2,75 2,50 2,25 2,00 1,75 1,50 1,25 1,00 0,75 0,50 0,25 0,00 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Number of Clusters

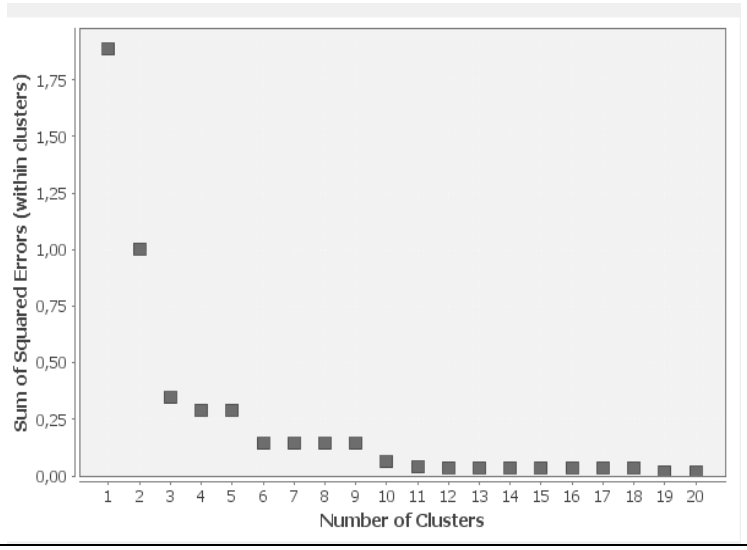
B1	**2**	

Fig.17. Proceso Elbow

En las primeras 100 paginas, código B5, se ve en el grafico que también existirían posibles valores de Temas en 3 y 4, dado que la suma de errores al cuadrado da exactamente igual, el programa arroja el óptimo de Temas en 2, por lo que se considera que es desde allí se vuelve un valor marginal.

Cabe recordar que este método busca una suma de cuadrados pequeña, sin tener una tendencia a cero del error, es donde hay un decenso considerable del error con respecto a los Clúster con lo que se determina el valor optimo, gráficamente se toma como si el codo se doblara de un brazo. Después de esto, se aplica el nodo LDA y los resultados que se obtienen son los siguientes Topics:

	B5	**B4**	**B3**	**B2**	**B1**
	Dela	Asi	Dela	Mrs	Ciudad
	Dicha	Cabillo	Dichos	Dichos	Casas
	cabillo	Pena	Cabillo	Año	Yglesia
	Casas	Dean	Dean	Cabillo	Cabillo
TOPIC 0	Año	Casa	Mrs	Bienes	Plasencia

	Mas Renta Mrs Dia san	Casas Viña Muger Martínez	Yglesia Casas Juan Dia é	Vida Pena Renta Así Non	Gonzalez Cabildo Notario Gutiérrez Dha
Topic 1	Dicha Dela Dichos Cabillo Yglesia Mrs Pedro Notario Gonzalez qual	Cabillo Yglesia Renta Plasencia Fijo Beneficiados Muger Ciudad Fija Ano	Dela Camino Mojon Ba Bá Tierra Dende Diego Dá Cerro	Gonzalez Juan García Fernández Yglesia Notario Pedro Rui Plasencia Ciudad	Año Vos Mrs Pena Asi Renta Non Dia Delos gallinas
Topic 2				Dichos Cabillo Ciudad Casas Dean Cabildo Yglesia Dha Asi delos	

Contrastando estos resultados con la historia de Plasencia medieval se pueden obtener las siguientes, generalizaciones de Temas tratados dentro del documento, cabe aclarar que se es necesario acudir a documentos históricos para tener claridad del rol de algunos personajes

que aquí se mencionan, se analiza cada una de la relación de palabras para así establecer el tema del cual se está tratando.

	Topic 0	Topic 1	Topic 2
B5	**RENTA:** Las transacciones de renta de bienes se hacían en el cabildo.	**EXENTOS:** En 1255 el rey Alfonso Xinicio una política dónde los caballeros quedaban exentos de tributos, Pedro Sánchez, caballero al servicio de Alfonso X. Lora Serrano, G. (2001).	
B4	**PROCURADOR:** Miguel Martínez es procurador en 1325 encargándose de dar dahesa – terreno extenso generalmente acotado y dedicado al pasto del ganado - para los bueyes, delimitandola y asignándola. (Lora Serrano, 2001)	**BENEFICIADOS:** El cabildo es el gobierno local de la época, dónde se cobraban impuesto o se eximían de ello, se presume es un beneficio a las mujeres.	
	REGIDORES: Juan Fernández y Miguel Sánchez, son los regidores - gobernantes o concejales - y escribanos del rey	**DELIMITACIÓN DE PREDIOS.** Por su ubicación estratégica y excelente calidad de sus pastos, Plasencia serbia de paso temporal de ganado, este tránsito de mercancía y	

B3	Alfonso XI , Sánchez, A., Sánchez, A., & perfil, V. (2017)	ganado generaba gravámenes para la ciudad, El significado de mojón - Poste de piedra o cualquier señal clavada en el suelo que sirve para marcar el límite de un territorio o de una propiedad- presuntamente Diego Martínez quien cumpliría el papel de regidor Canalejo, E. (1983).	
B2	**RENTA:** Las transacciones de renta de bienes se hacían en el cabildo.	**CAMBIO DE REINADO:** Juan Fernández es el condestable (actúa en nombre del rey), cierra en concejo por orden de Enrique II, en el traspaso de reinado, cuando muere Pedro I, es el momento de saquear al conde García de Toledo y envía a los regidores uno de ellos Gutiérrez González. Sánchez, A., Sánchez, A., & perfil, V. (2017).	**BIENES:** Registro de propiedades a iglesia y cabildo.
	DISPUTA:		

B1	1428 una de las actividades que estaba prohibida en Plasencia, era el transporte de vino de otras partes de España, para comercializar, el concejo de Plasencia se percató que la iglesia estaba comercializando vino aprovechando su poder, evitando pagar así impuestos, los concejales, ¨ El Dr. Garci López de Carvajal, Gutiérrez González de Trejo y Alfonso Fernández de Logroño revisaron los cargos de indecencia y el abuso de los privilegios de la iglesia¨ por lo que ordenaron a través del notario ¨.., desde este punto en adelante, todo el vino que entra en la ciudad sólo pasará a través de esta puerta y no otro. Todos los otros productos, si son animales o recipientes, se pueden tomar a través de cualquier otra puerta¨. ("Churchmen and Their Illicit Wine Dealings, circa 1428 – Revealing Cooperation and Conflict Project", 2017)	**EXENTOS:** Algunos animales y productos fueron exentos del pago de impuestos (Lora Serrano, 2001)	

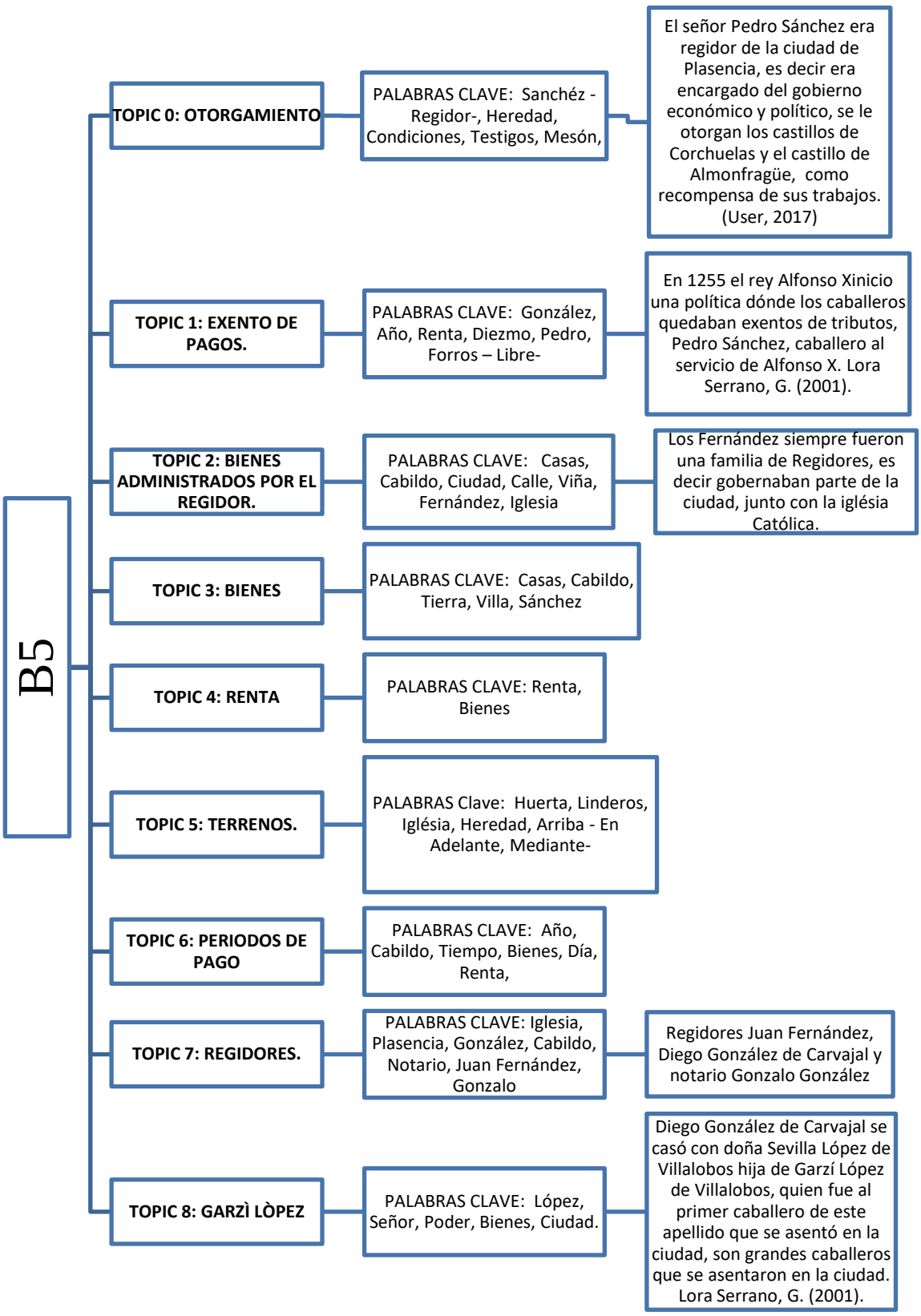
En los anexos se podrá encontrar las tablas originales arrojadas por el programa.

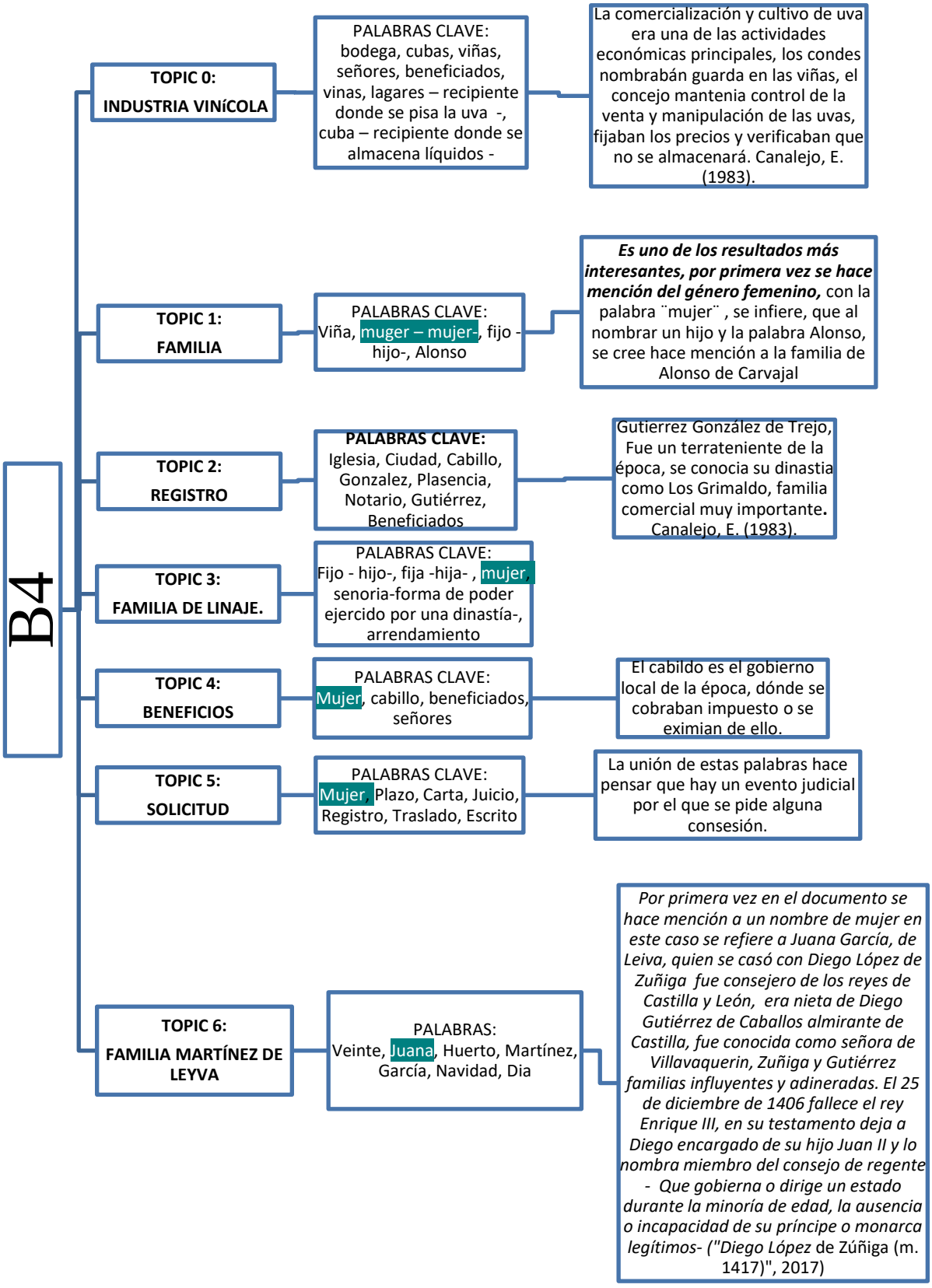
En la investigación, al implementar el método de obtención de tópicos Elbow, se evidencian ciertas relaciones interesantes en la investigación, por lo que en aras de aprovechar el máximo

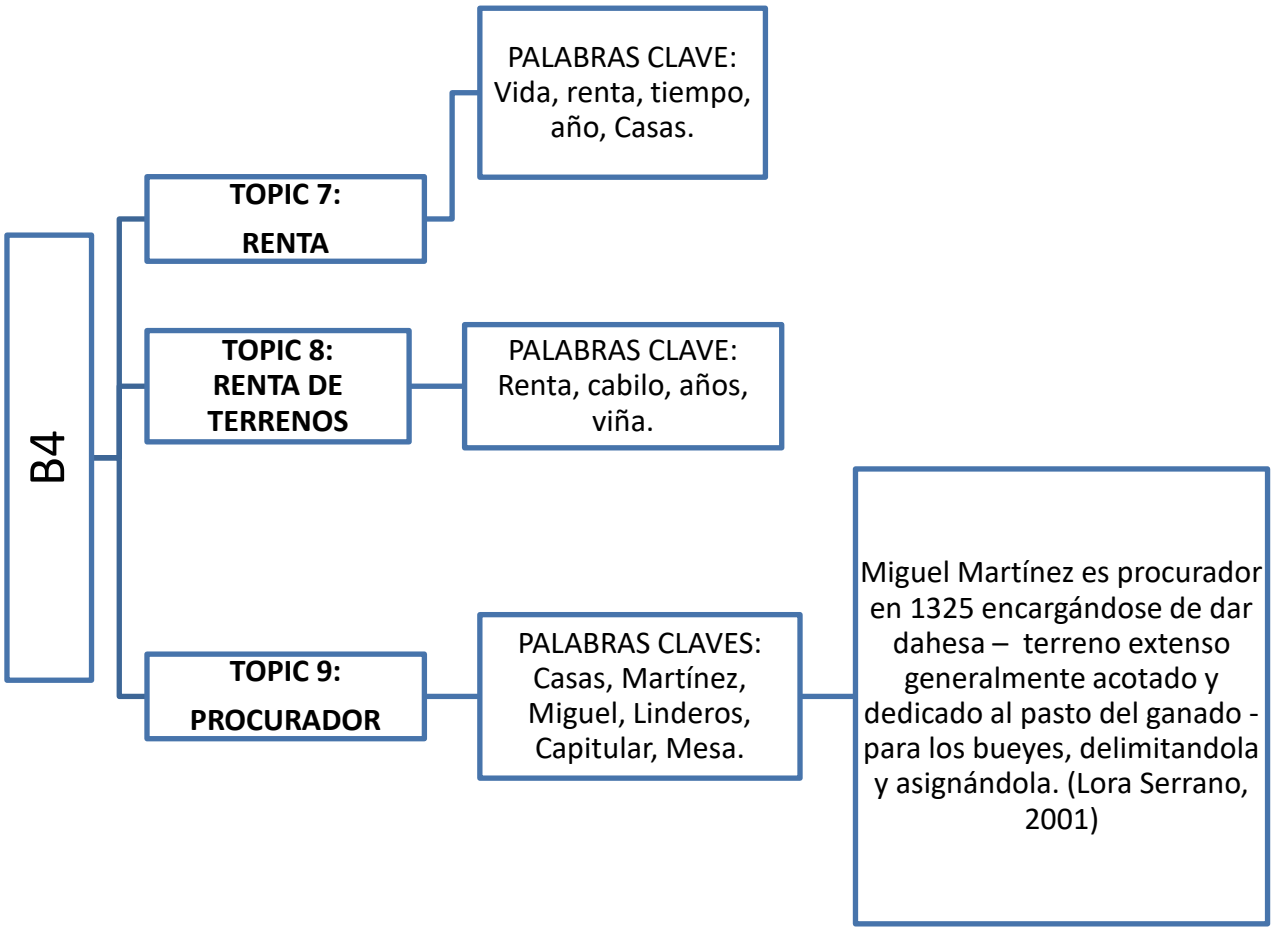
el proceso de extracción de topicos, que se logró desarrollar y en búsqueda de ampliar los temas, se realiza un ejercicio mostrando los datos arrojados de de 10 Temas, con 10 palabras, un se toma como criterio el máximo de probabilidades por palabra y al establecer 10 como tema y 10 como palabras, se está garantizando trabajar con las mayores probabilidades ligadas a los temas, la extracción de información es uno de los propósitos del ejercicio.

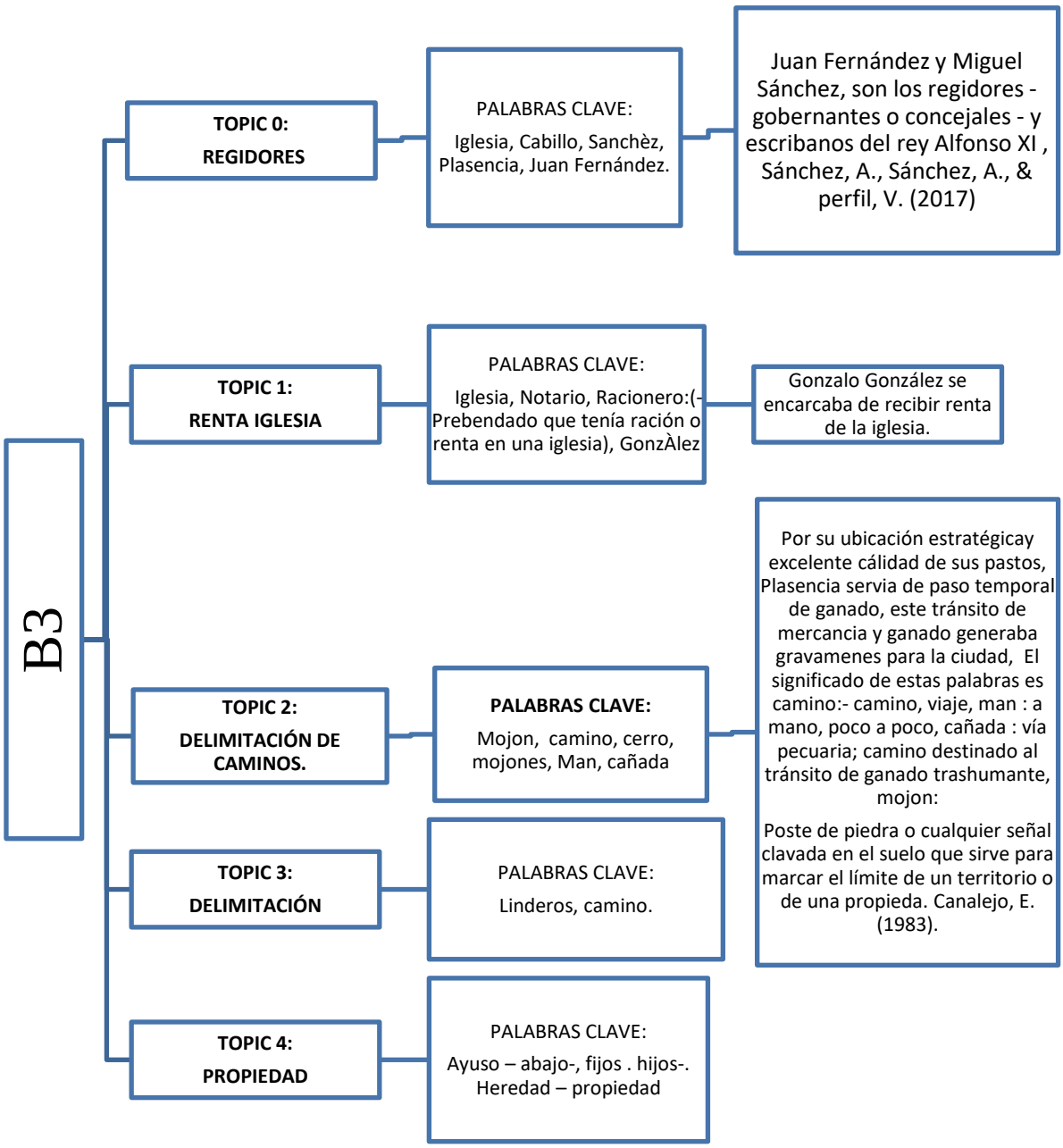
A continuación, se muestran los resultados arrojados de este ejercicio, se presenta en forma de mapa conceptual, relacionando el Tema que une las palabras obtenidas, con el análisis de las mismas, este análisis se hace a la luz de documentos propios de la historia de Plasencia, a pesar de que nuestra investigación no es histórica, pero esto ayuda a mejorar el entendimiento de los resultados obtenidos.

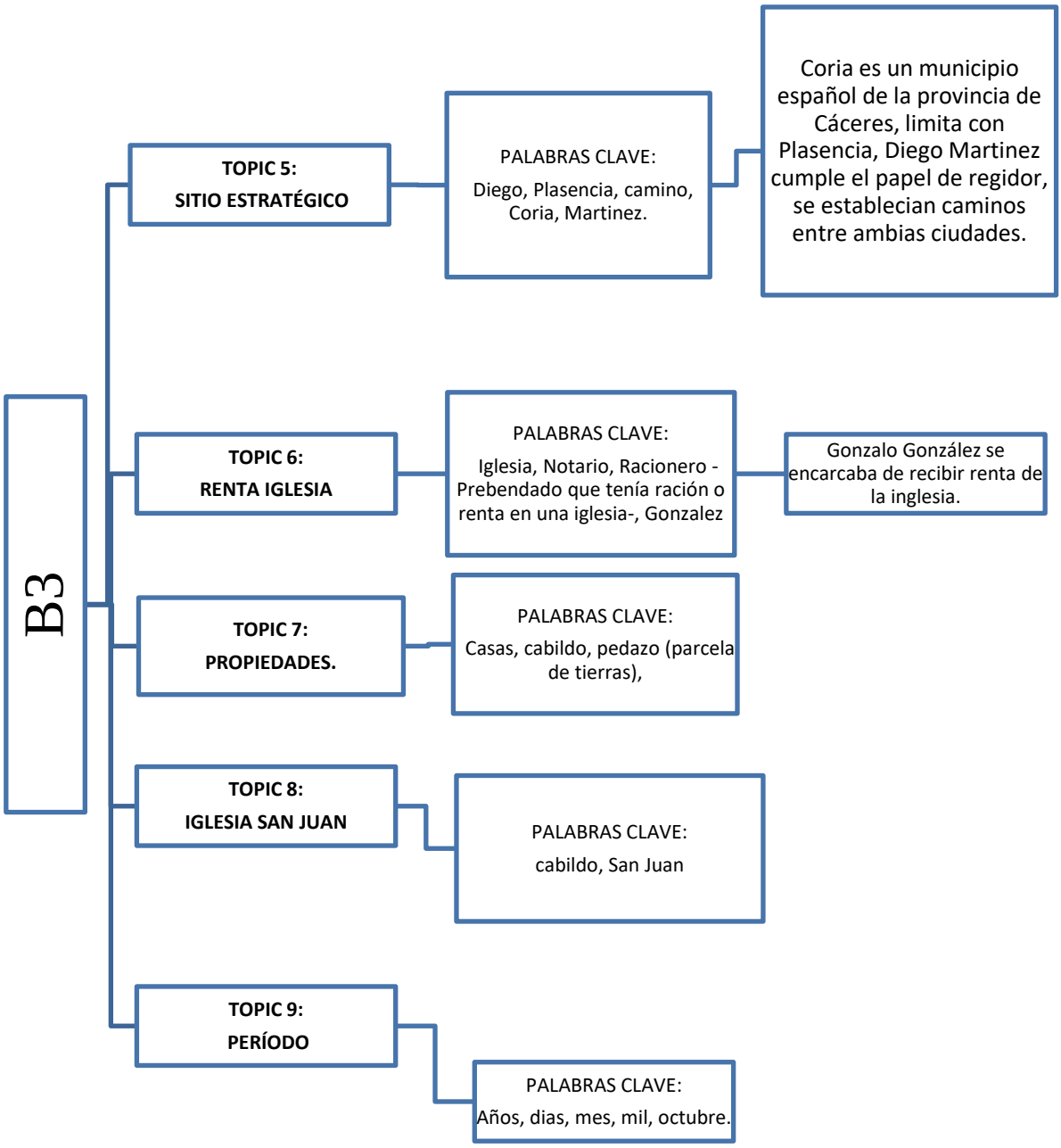
4.2.Método heurístico 10 temas, 10 palabras.

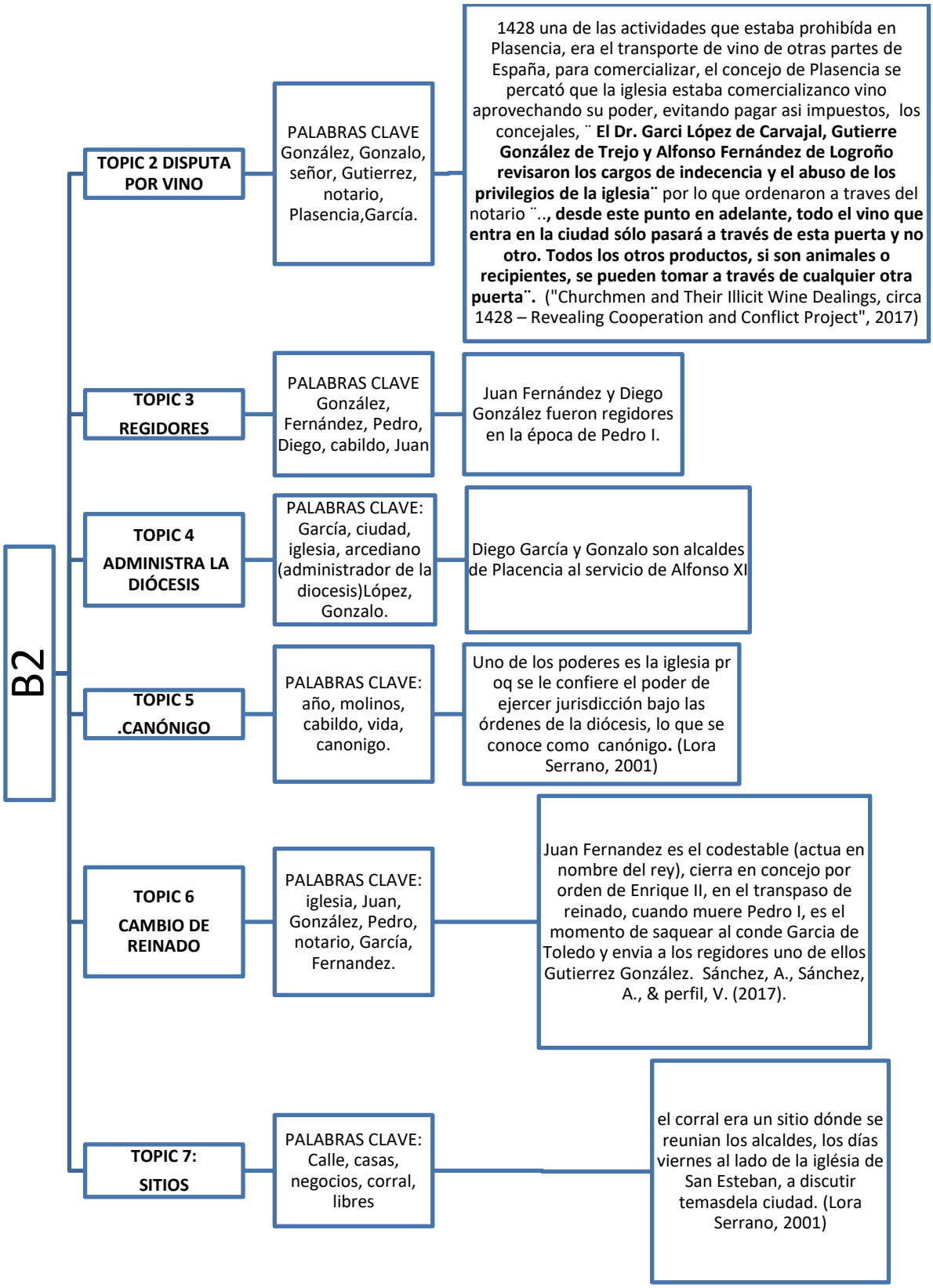


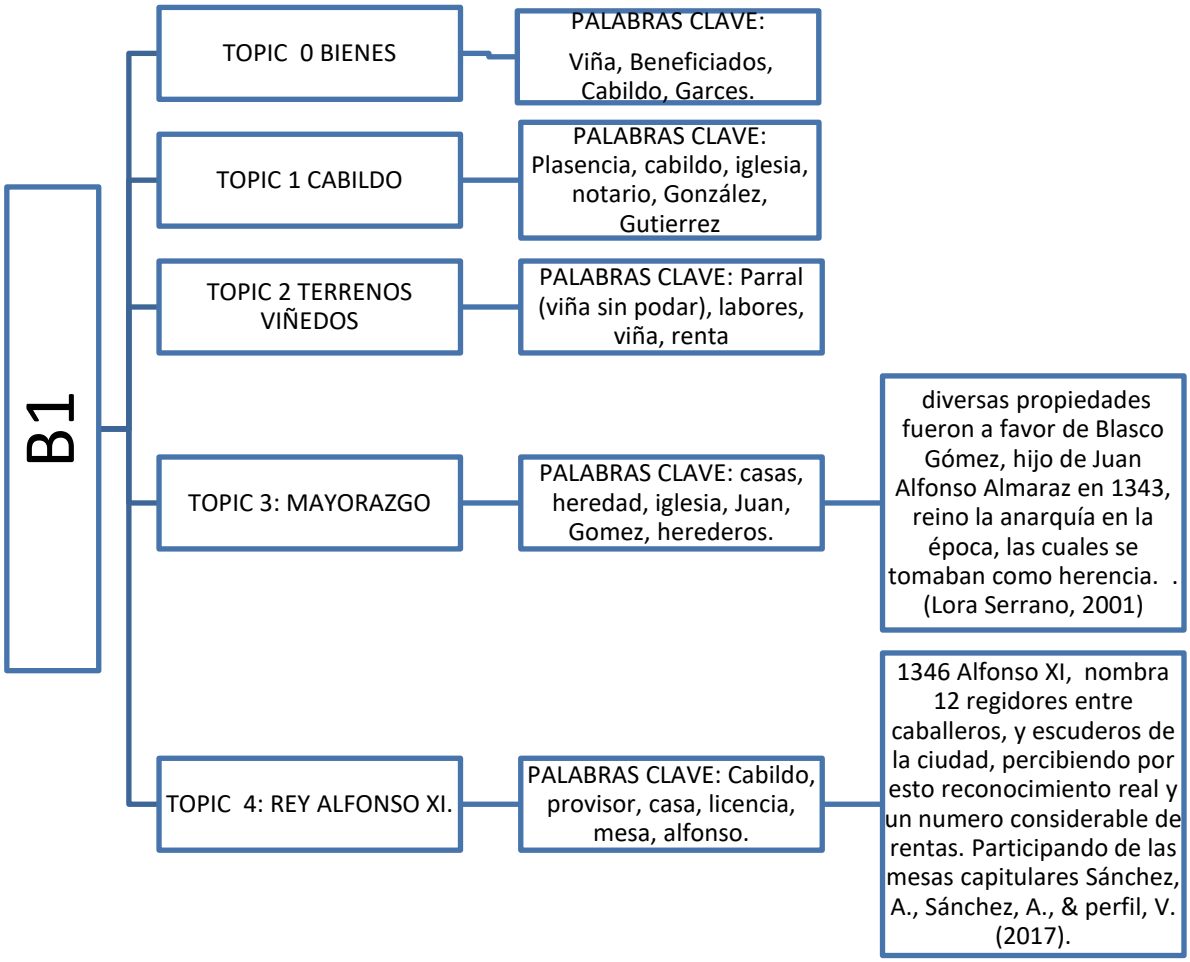

B4
TOPIC 0: INDUSTRIA VINíCOLA
PALABRAS CLAVE: bodega, cubas, viñas, señores, beneficiados, vinas, lagares – recipiente donde se pisa la uva -, cuba – recipiente donde se almacena líquidos -
La comercialización y cultivo de uva era una de las actividades económicas principales, los condes nombrabán guarda en las viñas, el concejo mantenia control de la venta y manipulación de las uvas, fijaban los precios y verificaban que no se almacenará. Canalejo, E. (1983).
TOPIC 1: FAMILIA
PALABRAS CLAVE: Viña, muger – mujer-, fijo - hijo-, Alonso
Es uno de los resultados más interesantes, por primera vez se hace mención del género femenino, con la palabra ¨mujer¨ , se infiere, que al nombrar un hijo y la palabra Alonso, se cree hace mención a la familia de Alonso de Carvajal
TOPIC 2: REGISTRO
PALABRAS CLAVE: Iglesia, Ciudad, Cabillo, Gonzalez, Plasencia, Notario, Gutiérrez, Beneficiados
Gutierrez González de Trejo, Fue un terrateniente de la época, se conocia su dinastia como Los Grimaldo, familia comercial muy importante. Canalejo, E. (1983).
TOPIC 3: FAMILIA DE LINAJE.
PALABRAS CLAVE: Fijo - hijo-, fija -hija- , mujer, senoria-forma de poder ejercido por una dinastía-, arrendamiento
TOPIC 4: BENEFICIOS
PALABRAS CLAVE: Mujer, cabillo, beneficiados, señores
El cabildo es el gobierno local de la época, dónde se cobraban impuesto o se eximian de ello.
TOPIC 5: SOLICITUD
PALABRAS CLAVE: Mujer, Plazo, Carta, Juicio, Registro, Traslado, Escrito
La unión de estas palabras hace pensar que hay un evento judicial por el que se pide alguna consesión.
TOPIC 6: FAMILIA MARTÍNEZ DE LEYVA
PALABRAS: Veinte, Juana, Huerto, Martínez, García, Navidad, Dia
Por primera vez en el documento se hace mención a un nombre de mujer en este caso se refiere a Juana García, de Leiva, quien se casó con Diego López de Zuñiga fue consejero de los reyes de Castilla y León, era nieta de Diego Gutiérrez de Caballos almirante de Castilla, fue conocida como señora de Villavaquerin, Zuñiga y Gutiérrez familias influyentes y adineradas. El 25 de diciembre de 1406 fallece el rey Enrique III, en su testamento deja a Diego encargado de su hijo Juan II y lo nombra miembro del consejo de regente - Que gobierna o dirige un estado durante la minoría de edad, la ausencia o incapacidad de su príncipe o monarca legítimos- ("Diego López de Zúñiga (m. 1417)", 2017)


B4
TOPIC 7: RENTA
PALABRAS CLAVE: Vida, renta, tiempo, año, Casas.
TOPIC 8: RENTA DE TERRENOS
PALABRAS CLAVE: Renta, cabilo, años, viña.
TOPIC 9: PROCURADOR
PALABRAS CLAVES: Casas, Martínez, Miguel, Linderos, Capitular, Mesa.
Miguel Martínez es procurador en 1325 encargándose de dar dahesa – terreno extenso generalmente acotado y dedicado al pasto del ganado - para los bueyes, delimitandola y asignándola. (Lora Serrano, 2001)


B3
TOPIC 0:
REGIDORES
PALABRAS CLAVE:
Iglesia, Cabillo, Sanchèz, Plasencia, Juan Fernández.
Juan Fernández y Miguel Sánchez, son los regidores - gobernantes o concejales - y escribanos del rey Alfonso XI , Sánchez, A., Sánchez, A., & perfil, V. (2017)
TOPIC 1:
RENTA IGLESIA
PALABRAS CLAVE:
Iglesia, Notario, Racionero:(-Prebendado que tenía ración o renta en una iglesia), GonzÀlez
Gonzalo González se encarcaba de recibir renta de la iglesia.
TOPIC 2:
DELIMITACIÓN DE CAMINOS.
PALABRAS CLAVE:
Mojon, camino, cerro, mojones, Man, cañada
Por su ubicación estratégicay excelente cálidad de sus pastos, Plasencia servia de paso temporal de ganado, este tránsito de mercancia y ganado generaba gravamenes para la ciudad, El significado de estas palabras es camino:- camino, viaje, man : a mano, poco a poco, cañada : vía pecuaria; camino destinado al tránsito de ganado trashumante, mojon:
Poste de piedra o cualquier señal clavada en el suelo que sirve para marcar el límite de un territorio o de una propieda. Canalejo, E. (1983).
TOPIC 3:
DELIMITACIÓN
PALABRAS CLAVE:
Linderos, camino.
TOPIC 4:
PROPIEDAD
PALABRAS CLAVE:
Ayuso – abajo-, fijos . hijos-. Heredad – propiedad


B3
TOPIC 5:
SITIO ESTRATÉGICO
PALABRAS CLAVE:
Diego, Plasencia, camino, Coria, Martinez.
Coria es un municipio español de la provincia de Cáceres, limita con Plasencia, Diego Martinez cumple el papel de regidor, se establecian caminos entre ambias ciudades.
TOPIC 6:
RENTA IGLESIA
PALABRAS CLAVE:
Iglesia, Notario, Racionero - Prebendado que tenía ración o renta en una iglesia-, Gonzalez
Gonzalo González se encarcaba de recibir renta de la inglesia.
TOPIC 7:
PROPIEDADES.
PALABRAS CLAVE:
Casas, cabildo, pedazo (parcela de tierras),
TOPIC 8:
IGLESIA SAN JUAN
PALABRAS CLAVE:
cabildo, San Juan
TOPIC 9:
PERÍODO
PALABRAS CLAVE:
Años, dias, mes, mil, octubre.


B2
TOPIC 2 DISPUTA POR VINO
PALABRAS CLAVE González, Gonzalo, señor, Gutierrez, notario, Plasencia,García.
1428 una de las actividades que estaba prohibída en Plasencia, era el transporte de vino de otras partes de España, para comercializar, el concejo de Plasencia se percató que la iglesia estaba comercializanco vino aprovechando su poder, evitando pagar asi impuestos, los concejales, ¨ El Dr. Garci López de Carvajal, Gutierre González de Trejo y Alfonso Fernández de Logroño revisaron los cargos de indecencia y el abuso de los privilegios de la iglesia¨ por lo que ordenaron a traves del notario ¨.., desde este punto en adelante, todo el vino que entra en la ciudad sólo pasará a través de esta puerta y no otro. Todos los otros productos, si son animales o recipientes, se pueden tomar a través de cualquier otra puerta¨. ("Churchmen and Their Illicit Wine Dealings, circa 1428 – Revealing Cooperation and Conflict Project", 2017)
TOPIC 3 REGIDORES
PALABRAS CLAVE González, Fernández, Pedro, Diego, cabildo, Juan
Juan Fernández y Diego González fueron regidores en la época de Pedro I.
TOPIC 4 ADMINISTRA LA DIÓCESIS
PALABRAS CLAVE: García, ciudad, iglesia, arcediano (administrador de la diocesis)López, Gonzalo.
Diego García y Gonzalo son alcaldes de Placencia al servicio de Alfonso XI
TOPIC 5 .CANÓNIGO
PALABRAS CLAVE: año, molinos, cabildo, vida, canonigo.
Uno de los poderes es la iglesia pr oq se le confiere el poder de ejercer jurisdicción bajo las órdenes de la diócesis, lo que se conoce como canónigo. (Lora Serrano, 2001)
TOPIC 6 CAMBIO DE REINADO
PALABRAS CLAVE: iglesia, Juan, González, Pedro, notario, García, Fernandez.
Juan Fernandez es el codestable (actua en nombre del rey), cierra en concejo por orden de Enrique II, en el transpaso de reinado, cuando muere Pedro I, es el momento de saquear al conde Garcia de Toledo y envia a los regidores uno de ellos Gutierrez González. Sánchez, A., Sánchez, A., & perfil, V. (2017).
TOPIC 7: SITIOS
PALABRAS CLAVE: Calle, casas, negocios, corral, libres
el corral era un sitio dónde se reunian los alcaldes, los días viernes al lado de la iglésia de San Esteban, a discutir temasdela ciudad. (Lora Serrano, 2001)


B1
TOPIC 0 BIENES
PALABRAS CLAVE:
Viña, Beneficiados, Cabildo, Garces.
TOPIC 1 CABILDO
PALABRAS CLAVE: Plasencia, cabildo, iglesia, notario, González, Gutierrez
TOPIC 2 TERRENOS VIÑEDOS
PALABRAS CLAVE: Parral (viña sin podar), labores, viña, renta
TOPIC 3: MAYORAZGO
PALABRAS CLAVE: casas, heredad, iglesia, Juan, Gomez, herederos.
diversas propiedades fueron a favor de Blasco Gómez, hijo de Juan Alfonso Almaraz en 1343, reino la anarquía en la época, las cuales se tomaban como herencia. . (Lora Serrano, 2001)
TOPIC 4: REY ALFONSO XI.
PALABRAS CLAVE: Cabildo, provisor, casa, licencia, mesa, alfonso.
1346 Alfonso XI, nombra 12 regidores entre caballeros, y escuderos de la ciudad, percibiendo por esto reconocimiento real y un numero considerable de rentas. Participando de las mesas capitulares Sánchez, A., Sánchez, A., & perfil, V. (2017).

5. DISCUSIÓN.

5.1.DISCUSIÓN MINERÍA DE TEXTO Y ANÁLISIS HISTÓRICO DE TEXTOS.

Sin duda la minería de texto es un aliado a la hora de hacer análisis de documentos, en este caso los documentos históricos constituyen datos no estructurados de los cuales es difícil hacer un análisis profundo con los métodos convencionales que aplican los historiadores. Sin embargo, en este trabajo encontramos que el trabajar con transcripciones hechas desde el español antiguo forzó a buscar alternativas para la detección de caracteres y de palabras que no se encuentran en diccionarios de palabras convencionales en Minería de texto. Por lo que el pre procesamiento de datos se tuvo que modificar en cuanto a una aplicación de los diccionarios habituales.

En general el proceso de etiquetado, limpieza y obtención final de tablas ¨limpias¨, es eficiente para la depuración de ruido dentro de los documentos originales. El uso de diccionarios para crear la bolsa de palabras a ser analizado es una de las limitantes que se encuentran en la mayoría de los programas predeterminados.

El método Elbow para la obtención de clúster, para la clasificación posterior de las tuplas de datos formadas después del procesamiento, es una buena alternativa cuando los datos no están clasificados como en este caso, sin embargo, no se debe descartar que la escogencia de numero de temas y de palabras por tema, también es un aspecto que tiene que analizar el investigador, ya que como se vio en este caso al tener mayor número de temas, se logra mayor número de relaciones ocultas, lo cual permite hacer mayor cantidad de inferencias y análisis de estos datos. Se evidencia que en los mapas conceptuales hay mayor cohesión de palabras que llevan a establecer de forma más natural el Tema del cual se está tratando.

La implementación de Latent Dirichlet Allocation, arroja resultados óptimos rápidos en donde no se detectan problemas de dimensionalidad, logrando una correcta asociación de palabras con temas y con documentos, el flujo que se construyó en Knime permite una programación practica y rápida.

5.2.DISCUSIÓN EL ROL DE LA MUJER EN PLASENCIA Y LOS DATOS HISTÓRICOS.

Las mujeres a lo largo de la historia han estado bajo la sombra de las labores y acciones de los hombres, en la edad media la mujer tenía diferentes dimensiones en la sociedad, dimensión familiar, posición social en la nobleza, ama de casa. Las dinámicas sociales dentro de las dinastías muestran mujeres del medioevo que se casaban por prolongar apellidos, por limar asperezas entre familias. Las que tenían algún tipo de linaje podían hacer transacciones comerciales como escrituración con el beneplácito de sus esposos, únicamente, (Estudios Medievales Hispánicos 5, 2016).

Sin embargo, a pesar de este conocimiento en los anales de la historia es importante verificar mediante estudios de minería si existen otros indicios que puedan llegar a dar luces de la forma de vida, en los resultados que se han encontrado por el método Elbow se llega a una mención de la mujer en la sección B4, las palabras que de allí se unen hablan de beneficios, iglesia y cabildo, pero no llega a desvelar del todo un rol. Al profundizar en 10 temas 10 palabras de asociación se muestra, la palabra hijo y la palabra Alonso, se cree hace mención de la familia de Alonso de Carvajal, también se hace referencia a ¨dinastía¨ como palabra asociada a mujer, así que la mujer de la cual se hace mención debe ser de linaje y como ya se mencionó, si están en las actas debió haber sido algún tipo de transacción comercial con una familia de abolengo.

Después se observa que el nombre Juana hace aparición, acompañado de dos apellidos Martínez y García, Por lo que se presume, se habla de Juana García de Leiva, quien se casó con Diego López de Zuñiga consejero de los reyes de Castilla y León, era nieta de Diego Gutiérrez de Caballos almirante de Castilla, fue conocida como señora de Villavaquerin, Zuñiga y Gutiérrez familias influyentes y adineradas. El 25 de diciembre de 1406 fallece el rey Enrique III, en su testamento deja a Diego encargado de su hijo Juan II y lo nombra miembro del consejo de regente - Que gobierna o dirige un estado durante la minoría de edad, la ausencia o incapacidad de su príncipe o monarca legítimos- ("Diego López de Zúñiga (m. 1417)", 2017)

Aunque los resultados que se obtienen no son concluyentes, para una caracterización profunda si deja en claro que las mujeres dentro de los documentos de la catedral de Plasencia son mujeres que provienen de familias reconocidas de la época.

Los estudios al respecto son nulos como lo afirma M. Nash y M. Ferrandis, (1991), el estudio del papel de las mujeres en el medioevo es mínimo y los escritos que existen son realizados por varones por lo que no se es claro el papel desempeñado.

6. CONCLUSIONES.

- El desarrollo de un proceso eficiente, desarrollado en un lenguaje amigable para historiadores, brinda una herramienta práctica para el estudio de textos históricos en español antiguo, a través de métodos de minería textual, en donde permite que el historiador realice un análisis concreto y rápido de documentos de su interés, sin implementar las formas tradicionales desde la historia, de análisis de textos.
- El proceso de Asignación latente de Dirichlet para **datos históricos,** muestra su eficacia al optimizar el tiempo de clasificación y análisis de textos, ya que permite en poco tiempo obtener los temas principales y las palabras que se ligan a estos.
- Referente a el objetivo ejemplo principal usado en el prototipo de la investigación se concluye que, en *las actas capitulares,* la mención de las mujeres se hace únicamente en el código B4, y esta hace referencia a ***la mujer como miembro de familia, encargada de velar por los hijos o esposa de un hombre influyente***, más no, como parte de las actividades económicas que se registran en las actas. Por lo que podemos concluir que las actividades de las cuales se daba cuenta, en las actas únicamente hacían mención exclusiva a los miembros de la iglesia y del concejo de la ciudad. La información obtenida nos es insuficiente para caracterizar muchos más aspectos de la mujer en Plasencia, sin embargo, esto es determinante para así concluir que la participación en ámbitos económicos era exclusiva a mujeres de linaje, como se mostró en la discusión de resultados.
- Los temas principales dentro de las actas capitulares, que se evidencian, en los *grupos no observados arrojados por el proceso propuesto con el modelo* LDA, muestran:
    - ***Gobernantes*** *es recurrente la mención de regidores y concejales de la ciudad.*
    - ***Iglesia:*** *en todas las relaciones comérciales y familiares de Plasencia, la iglesia católica ha estado involucrada, de los resultados arrojados se muestra una relación con los recaudos de dineros por concepto de renta, siempre su aparición está constantemente relacionada con el cabildo, puede estimarse que esta institución es un gobernante más. Las familias adineradas y de dinastía siempre están ligadas a la iglesia. Además, la delimitación y control*

de terrenos, entre ellos la industria vinícola, también está relacionada con la iglesia.

- ***Industria vinícola:*** *la palabra viña, ligada a renta y terrenos es una constante en los resultados arrojados, se establece así que esta es una de las actividades económicas principales, la aparición de renta con terrenos establece que algunos de estas hectáreas implementadas en esta industria eran alquiladas y de las cuales el cabildo recibía dinero por concepto de renta.*
- ***Delimitación de terrenos:*** *el establecer límites y delimitación de terreno aparece frecuentemente unido con la palabra renta, el constante cercamiento hace pensar en disputas por el terreno o negocio del cabildo y la iglesia.*
- ***Apellidos y nombres:*** *los apellidos de hombres de familias adineradas y de grandes dinastías siempre están ligadas al cabildo, en muy pocas ocasiones se hace mención de nombres propios, son muy pocas la aparición de nombre y apellido, únicamente aparece un nombre de mujer Juana.*
- ***Bienes:*** *cabildo, casas iglesia, forman una composición que aparece con frecuencia lo que hace pensar que se llevaba un registro de las casas en el cabildo y en la iglesia.*
- ***Renta:*** *tal vez una de las palabras que más apareció en los resultados, evidencia que este modelo comercial hacia parte de la economía cotidiana, de los ciudadanos unos eran arrendadores (cabildo – iglesia) y otros los arrendatarios familias sin poder económico.*
- ***Periodos:*** *se habla de meses, días, años y en ocasiones alguna cantidad en letras,* ***siempre,*** *ligada a la renta, se cree eran los periodos de pago y tiempo establecidos de arrendamiento.*

❖ Para investigaciones a futuro se pude estudiar una división automática: por fechas, periodos, por familias influyentes de la época.

7. REFERENCIAS.

- Aggarwal C.C & Zhai C. (2012). *Mining Text Data*, United State: Springer
- Au Yeung, C. & Jatowt, A. (2011). "Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining". en *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York, USA. pp. 1231–1240. Recuperado de http://dl.acm.org/citation.cfm?id=2063755
- Alvarez Ramos, L. (2017). Revisado de https://riunet.upv.es/bitstream/handle/10251/89807/ALVAREZ%20-%20An%C3%A1lisis%20de%20ciudades%20a%20trav%C3%A9s%20su%20actividad%20en%20redes%20sociales.pdf?sequence=1, (consultado 7 de Noviembre, 2017)
- Antonelli, S., Castaño, J., Rocco, L., & Picoaga, J. (2013). *El discurso de la nación Argentina (2011- 2013), Un abordaje desde minería de texto* (Trabajo práctico final). Recuperado de: https://leonardosrocco.files.wordpress.com/2013/08/final-text-mining-senadores-antonelli_rocco.pdf
- Análisis de Conglomerados (2015). En *Unesco.org.* Recuperado de: http://www.unesco.org/webworld/portal/idams/html/spanish/S2CLUSFI.htm (Consultado en octubre 22, 2015).
- Árboles de Decisión. (n.d.). En *Monografías*. Recuperado de: http://www.monografias.com/trabajos56/mineria-de-datosvenezuela/mineria-de-datos-venezuela.shtml. (Consultado en octubre 13, 2015)
- Anon. (2015). En *Grupo MINA*. Recuperado de: http://www.fing.edu.uy/inco/cursos/fpr/wiki/index.php/Tuplascualitativos (Consultado en octubre 13,. 2015).
- Berry, M. & Kogan, J. (2010). *Text mining* (1st ed.). Hoboken, New Jersey: John Wiley & Sons.
- Blei, D., Y.Ng, A. & Jordan, M. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research,* [online] 3, pp.993-1022. Recuperado de: http://jmlr.csail.mit.edu/papers/v3/blei03a.html (Consultado en mayo 29, 2017)

- Buechler, M., Heyer, G. & Gründer, S. (2008). "eAQUA–bringing modern text mining approaches to two thousand years old ancient texts". In *Proceedings of e-Humanities–An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.
- Canalejo, E. (1983). *La vida económica de Plasencia en el siglo XV. Vol III,* Revistas.ucm.es. (Consultado Julio 2, 2017) http://revistas.ucm.es/index.php/ELEM/article/view/ELEM8282220553A/25267
- Ciula, A., Spence, P. & Vieira, J. (2008). *Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project.* London, UK: Centre for Computing in The Humanities, King´s College London.
- Comentario de textos. (2017). Claseshistoria.com. Recuperado el 22 Octubre de 2017, de http://www.claseshistoria.com/general/comentariotextos.htm
- Churchmen and Their Illicit Wine Dealings, circa 1428 – Revealing Cooperation and Conflict Project. (2017). Revealingcooperationandconflict.com. Retrieved 3 July 2017, from http://revealingcooperationandconflict.com/churchmen-and-their-illicit-wine-dealings-circa-1428/
- Concepto de Corpus y su definición. (2015). En *Estudios de Lingüística en Español*, online. Recuperado de: http://elies.rediris.es/elies18/23.html (Consultado en octubre 14, 2015).
- Daniel T. Larose. Discovering Knowledge in Data: An Introduction to Data Mining, USA, (2014),
- Ferrer, J. (2010). *Conceptos básicos de la metodología de la investigación*. Metodología de la investigación. Recuperado de: metodologia02.blogspot.com.co
- Francis,L., FCAS, MAAA, & Flynn, M. (2010). "Text Mining Handbook". En *Lou Casualty Actuarial Society E-Forum*, Spring. Recuperado de: https://www.casact.org/pubs/forum/10spforum/Francis_Flynn.pdf
- Guyo, I. (2008). Introduction To Machine Learning.
- Gove, R. (2017). Using the elbow method to determine the optimal number of clusters for k-means clustering. Bl.ocks.org. Revisado de

https://bl.ocks.org/rpgove/0060ff3b656618e9136b, (Consultado Noviembre 7, 2017)

- KNIME. (2014). Text mining workflow. Recuperado de: http://www.dataminingreporting.com/uploads/4/0/9/7/4097240/text_mining_tutorial_knime_sep_2014.pdf [consultado en junio 2, 2017]
- KNIMETV. (2013, 11, 01). Text Mining Webinar [Archivo de vídeo]. Recuperado de: https://www.youtube.com/watch?v=tY7vpTLYlIg&list=PLz3mQ6OlTI0ZQh-kH-ckZqngJU0HbKNjP&index=1 [Consultado en octubre 14, 2015].
- Kimura, F., Osaki, T., Tezuka, T., & Maeda, A. (2013). "Visualization of relationships among historical persons from Japanese historical documents". *Literary and linguistic computing*, 28(2), pp. 271-278
- KNIME Open Source Story | KNIME. (2017). Knime.com. https://www.knime.com/knime-open-source-story (Consultado en November 5, 2017)
- Latent Dirichlet Allocation. (2017). Es.wikipedia.org. recuperado, 4 July 2017, from https://es.wikipedia.org/wiki/Latent_Dirichlet_Allocation
- Latent Dirichlet Allocation. (2017). En *Msdn.microsoft.com*. Recuperado de: https://msdn.microsoft.com/en-us/library/mt762914.aspx (Consultado en Junio 5, 2017).
- Lora Serrano, G. (2001). *El primer gobierno municipal de Plasencia. from Recuperado de:* https://dialnet.unirioja.es/servlet/articulo?codigo=625179, *(Consultado Junio 11, 2017)*
- Molina López, J.M. & García Herrero, J. (2006). *Técnicas De Análisis De Datos Aplicaciones Prácticas Utilizando Microsoft, Excel y Weka*, Universidad Carlos III de Madrid. Recuperado de: http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf
- Martínez , R. (2016). *Decipherin secrets:Unlocking the manuscripts of medieval spain.* Curso Online, recuperado en agosto 4, 2016 de:

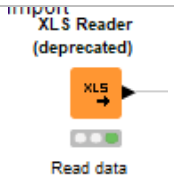
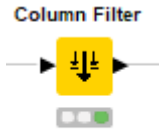
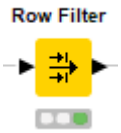
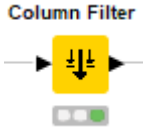
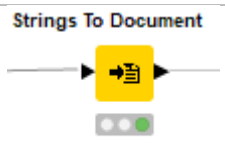
- *http://decipheringsecrets.net/themes/default/documents/Deciphering%20Secrets%20Paleography%20Manual%20Version%201.1.pdf.*
- Minería de Datos. (n.d.). En *Msdn.microsoft.com.* Recuperado de: https://msdn.microsoft.com/es-es/library/ms174949(v=sql.120).aspx (Consultado en agosto 10, 2015)
- Medoides. (n.d.). En *Unesco.org.* Recuperado de: http://www.unesco.org/webworld/portal/idams/html/spanish/S2CLUSFI.htm (Consultado en agosto 10, 2015)
- Moreiro González, J. (2002). Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información. Retoc.iula.upf.edu. de http://retoc.iula.upf.edu/html/n-gram.pdf (consultado en noviembre 26, 2017)
- Meyer C, et al. (2015) Informe fundación Bankinter, Big Data.
- McDonald, D. (2014). "A Text Mining Analysis of Religious Texts". En *The Journal of Business*, 13(1), pp. 27-47.
- M. Nash y M Ferrandis (1991), Dos décadas de historia de las mujeres en España: una reconsideración, Historia Social, No. 9 (Winter, 1991), pp. 137-161, disponible en http://www.jstor.org/stable/40340551.
- Probabilidad a priori. (2017). En *Enciclopedia libre Wikipedia online*. Recuperado de: https://es.wikipedia.org/wiki/Probabilidad_a_prior (Consultado en junio 5, 2017).
- Qahl, S. H. M. (2014). *An Automatic Similarity Detection Engine Between Sacred Texts Using Text Mining and Similarity Measures.* (Tesis doctoral), Rochester Institute of Technology, New York.
- Recto y verso. (2016). En *Enciclopedia libre Wikipedia online*. Recuperado de: https://es.wikipedia.org/wiki/Recto_y_verso (Consultado en diciembre 21, 2016)
- Sintaxis (2015). En Gramaticas.net. Recuperado de: http://www.gramaticas.net/2013/01/la-sintaxis.html (Consultado en octubre 14, 2015).

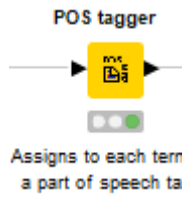
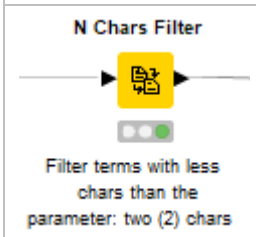
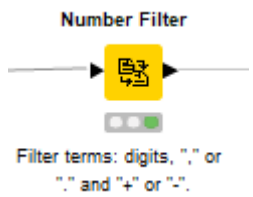
- Swanson, D.R., Smalhaiser, N.R. (1994). "Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease". En *Neuroscience research communications*. Vol. 15, pp.1-9.
- Socha Díaz, D., Martínez Serna, J., & Medina Mosquera, C. (2017). *Minería de texto histórica - colaboración al proyecto "Revealing Cooperation and Conflict Project"*. Repositorio.escuelaing.edu.co, Revisado de:http://repositorio.escuelaing.edu.co/handle/001/521Text Mining (2015).
- Semántica. (2015). EN *Definición.de online.* Recuperado de: http://definicion.de/semantica/ (Consultado octubre 17, 2015].
- Sánchez, A., Sánchez, A., & perfil, V. (2017). *"SIGLO XIV EN PLASENCIA: LA BAJA EDAD MEDIA". (PARTE II). Parragasanchezalfonso.blogspot.com.co. Revisado 1 April 2017, Revisado de http://parragasanchezalfonso.blogspot.com.co/2016/05/siglo-xiv-en-plasencia-la-baja-edad.html*
- En *Textmining.galeon.com*. Recuperado de: *http://textmining.galeon.com/#Concepto* (Consultado en octubre 13, 2015)
- Tf-idf. (2017). En *Enciclopedia libre Wikipedia online.* Recuperado de: https://es.wikipedia.org/wiki/Tf-idf (Consultado en Mayo 29, 2017).
- Topic Modeling. (2017). Mallet.cs.umass.edu. Revisado en http://mallet.cs.umass.edu/topics.php, (consultado Noviembre 7, 2017)
- Triantaphyllou, E. & Liao, T. (2008). "Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications". En *Singapore: World Scientific Publishing Company*. Vol. 6, University of Florida.
- The Stanford Natural Language Processing Group. (2017). Nlp.stanford.edu. Revisado de: https://nlp.stanford.edu/software/tagger.shtml, (Consultado Noviembre 6, 2017)
- User, S. (2017). Castillo de Monfragüe. Ayuntamiento de Torrejón el Rubio. Parque Nacional de Monfragüe. (Revisado, 2 June 2017) Recuperado de: http://www.torrejonelrubio.com/index.php/monumentos-historicos-de-torrejon/149-castillo-de-monfraguee

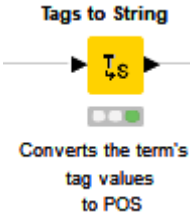
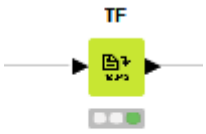
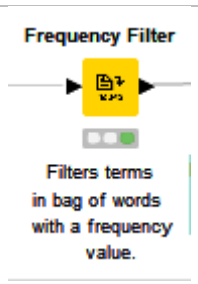
- Visa, A., Toivonen, J., & Hannu Vanharanta, B. B. (2002). “Contents matching defined by prototypes: Methodology verification with books of the Bible”. en *Journal of Management Information Systems*, 18(4), 87-100.
- Vallejos, S. (2006). Minería de datos (Trabajo de Adscripción). Universidad Nacional del Nordeste, Argentina.
- Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Second edition. San Francisco California: ELSEVIER.
- Xu, G-X., Qiu, L-R., Yang, L. (2014). "Tibetan text Clústering based on machine learning". En *The Free Library*, June 01. (Consultado en septiembre 9, 2015)

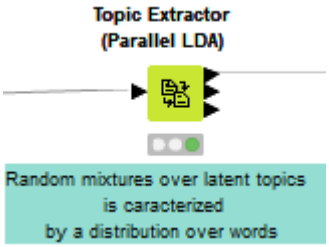
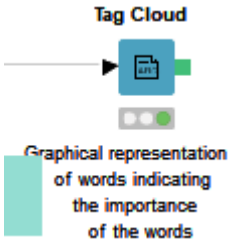
8. ANEXOS.

8.1. Nodos KNIME

	Nodo 1: XlS reader, permite la lectura de archivos desde Excel, los datos pueden ser numéricos o textuales.
	Nodo 2: Column Filter, se extraen los datos de las columnas, pasándolos a una nueva tabla de salida.
	Nodo 3: Sorter, organiza de las actas según el código B, en forma ascendente.
	Nodo 4: Row Filter, busca en las filas números o números ID.
	Nodo 5: GroupBy, borra las transcripciones que se encuentran repetidas.
	Nodo 6: Column Filter, se extraen los datos de las columnas, pasándolos a una nueva tabla de salida.
	Nodo 7: String To Document, convierte las filas en documentos.

	Nodo 8: POS Tagger, se etiqueta según POS Tagger - Part-Of-Speech Tagger (POS Tagger) esto permite hacer la asignación de una etiqueta como verbo, pronombre, nombres etcétera, a cada palabra.
	Nodo 9: Puntuaction Erause, remueve la puntuación del documento, filtra palabras con dos caracteres, tres no por las abreviaciones como la palabra don.
	Nodo 10: N chars Filter, quita todos aquellos caracteres como: +, -, x, /, palabras como (y/o)
	Nodo 11: number filter, filtra todos los números del documento.
	Nodo 12: Case convert, asigna nuevas columnas a la información ya filtrada.
	Nodo 13: Stop Word filter, filtran artículos y pronombres, obtenidos de los diccionarios de las actas.

	Nodo 14: Stop Word filter, filtran palabras consideradas vacías de las actas.
	Nodo 15: Bag of Words Creator, crea una bolsa de palabras las cuales constituirán las tuplas de información a ser analizas.
	Nodo 16: Tags to String, convierte el documento en cadenas, formando una nueva columna con las etiquetas del POS Tagger - Part-Of-Speech Tagger (POS Tagger), esto permite hacer la asignación de una etiqueta como verbo, pronombre, nombres etcétera, a cada palabra.
	Nodo 17: TF, calcula la frecuencia relativa de cada termino.
	Nodo 18: Frecuency Filter, filtra las palabras según unos límites impuestos, las de mayor frecuencia se mantienen.

	Numero de topicos
	Nodo 19: LDA, clasificador de las palabras según los temas encontrados según el método de Variables ocultas de Dirichlet.
	Nodo 20: Tag Cloud, por medio de un gráfico, según el tamaño y posición en texto, se muestran las palabras según su importancia.

8.2. Tablas originales arrojadas por el programa.

Row ID	Topic id	Term	Weight
Row0	topic_0	ciudad	958
Row1	topic_0	casas	658
Row2	topic_0	yglesia	546
Row3	topic_0	cabillo	530
Row4	topic_0	plasencia	485
Row5	topic_0	dha	460
Row6	topic_0	gonzalez	435
Row7	topic_0	cabildo	416
Row8	topic_0	notario	363
Row9	topic_0	gutierre	290
Row10	topic_1	año	920
Row11	topic_1	vos	874
Row12	topic_1	mrs	818
Row13	topic_1	pena	741
Row14	topic_1	asi	717
Row15	topic_1	renta	698
Row16	topic_1	non	571
Row17	topic_1	dia	444
Row18	topic_1	delos	414
Row19	topic_1	gallinas	394

Row ID	Topic id	Term	Weight
Row0	topic_0	dela	1,067
Row1	topic_0	dichos	933
Row2	topic_0	cabillo	888
Row3	topic_0	dean	763
Row4	topic_0	mrs	605
Row5	topic_0	yglesia	588
Row6	topic_0	casas	578
Row7	topic_0	juan	508
Row8	topic_0	dia	488
Row9	topic_0	é	462
Row10	topic_1	dela	1,140
Row11	topic_1	camino	978
Row12	topic_1	mojon	784
Row13	topic_1	bá	718
Row14	topic_1	ba	686
Row15	topic_1	tierra	646
Row16	topic_1	dende	581
Row17	topic_1	diego	581
Row18	topic_1	dá	574
Row19	topic_1	cerro	538

Row ID	Topic id	Term	Weight
Row0	topic_0	mrs	391
Row1	topic_0	dichos	351
Row2	topic_0	año	346
Row3	topic_0	cabillo	276
Row4	topic_0	bienes	229
Row5	topic_0	vida	219
Row6	topic_0	pena	217
Row7	topic_0	renta	214
Row8	topic_0	asi	214
Row9	topic_0	non	190
Row10	topic_1	gonzalez	601
Row11	topic_1	juan	563
Row12	topic_1	garcia	527
Row13	topic_1	ferrandez	483
Row14	topic_1	yglesia	471
Row15	topic_1	notario	386
Row16	topic_1	pedro	360
Row17	topic_1	rui	312
Row18	topic_1	plasencia	279
Row19	topic_1	ciudad	268
Row20	topic_2	dichos	757
Row21	topic_2	cabillo	579
Row22	topic_2	ciudad	528
Row23	topic_2	casas	495
Row24	topic_2	dean	416
Row25	topic_2	cabildo	401
Row26	topic_2	yglesia	389
Row27	topic_2	dha	347
Row28	topic_2	asi	335
Row29	topic_2	delos	294

Row ID	S Topic id	S Term	D Weight
Row0	topic_0	asi	882
Row1	topic_0	cabillo	657
Row2	topic_0	pena	652
Row3	topic_0	dean	491
Row4	topic_0	casa	447
Row5	topic_0	casas	424
Row6	topic_0	viña	414
Row7	topic_0	muger	362
Row8	topic_0	martinez	361
Row9	topic_0	año	359
Row10	topic_1	cabillo	741
Row11	topic_1	yglesia	614
Row12	topic_1	renta	565
Row13	topic_1	plasencia	550
Row14	topic_1	fijo	548
Row15	topic_1	beneficiados	544
Row16	topic_1	muger	508
Row17	topic_1	ciudad	502
Row18	topic_1	fija	479
Row19	topic_1	ano	478

Row ID	S Topic id	S Term	D Weight
Row0	topic_0	dela	951
Row1	topic_0	dicha	773
Row2	topic_0	cabillo	675
Row3	topic_0	casas	630
Row4	topic_0	año	626
Row5	topic_0	mas	486
Row6	topic_0	renta	424
Row7	topic_0	mrs	322
Row8	topic_0	dia	321
Row9	topic_0	san	318
Row10	topic_1	dicha	1,134
Row11	topic_1	dela	991
Row12	topic_1	dichos	896
Row13	topic_1	cabillo	664
Row14	topic_1	yglesia	595
Row15	topic_1	mrs	540
Row16	topic_1	pedro	500
Row17	topic_1	notario	383
Row18	topic_1	gonzalez	383
Row19	topic_1	qual	372